Why are things messed up?
Political economy basics*

James D. Fearon

DRAFT, Parts I-III

July 29, 2026

Part I

Introduction

1 Why are things messed up?

Maybe it is because people are often or always immoral or otherwise possessed of bad desires and beliefs. Sinful, in the Abrahamic religions, or prone to behaving incorrectly or holding incorrect beliefs in Confucianism and Marxist-Leninism, for example. Perhaps it is because the other political party and its supporters have immoral views or have been led astray by a bad ideology.

Another possibility is that things are messed up because people are not so much immoral as fallible. We try to do the right thing but we are less than fully rational. We suffer from weakness of the will and a panoply of biases and cognitive limitations that lead us to take actions that do not turn out as we would like.

This book develops a different kind of answer, which can be summarized by saying that often things are messed up because coordinating the actions of large numbers of people is just extremely difficult. By itself this statement is pretty unenlightening, but it becomes less so when we explore how and why large-scale coordination is so difficult in often-encountered *strategic situations* in politics and economics.

I do not think that either of the first two explanations is wrong or unimportant – to the contrary. But because the coordination problems can be so significant under reasonable assumptions about human moral and intellectual capacities, I expect that we will generally need to consider the strategic context when we propose as remedies moral education or interventions to mitigate the effects of biases and cognitive limitations.

What does "messed up" refer to? First off, things are less messed up the more we have efficient, effective, and equitable provision of a broad range of public services. For example: Garbage collection; clean water and air and protections against environmental degradation; basic security against violence and theft; transportation and other infrastructure; an accessible and impartial judicial system; appropriate regulation of pharmaceuticals, food safety, and other potentially hazardous goods; laws, policies, and administration that support public, private, or mixed social insurance for health, old age, and disability; available and accessible high quality education; a range of government policies and institutions that favor broad-based economic growth.

While there is relatively little disagreement about the value of these things, there is normally a lot of disagreement about who should pay for them, how to pay, and how much to buy. There is also a whole host of public policies over which people can have strongly conflicting views about what is best, and thus what would be the least messed-up policy. For example: tax rates; policies on gun control, immigration, abortion, zoning laws, religion, foreign policy, school curricula, and public provision of health insurance.

What "messed up" means in such cases differs depending on who you talk to, and the conflicts that arise are the stuff of politics. Still, I will argue that in many areas the following standard has much to recommend it: A policy is better the closer it is to what the average citizen would want if he or she were well informed about its costs and implications. In some policy areas the justification for such a standard may be utilitarian, in others not utilitarian but rather as a pragmatic compromise that reduces incentives for costly conflict that almost no one likes. More below (section 16). The basic idea is that things are often plausibly "messed up" when implemented policies are those desired by a small minority, to the detriment or dislike of a large majority of people, or are intensely disliked by a substantial minority against a majority that doesn't care that much.

2 Political failures

The last section is somewhat misleading. For instance, I will have nothing to say in what follows about the specifics of improving arrangements for garbage collection (which is *not* a trivial or unimportant collective action problem). Instead, the question of why are things so often so messed up will be addressed largely in regard to a set of common, higher-level political failures.

- Electoral democracy would seem to provide a simple solution. Voters can select capable politicians, and/or vote them out of office if they fail to provide efficient, effective, and equitable public services. They can vote out politicians who implement policies that are far from what typical voters want. The threat of losing office should motivate politicians to work to make things less messed up, and to choose broadly popular policies.

  But even when elections are quite competitive and fair, electoral democracy often fails to deliver. This is true in both high- and low-income countries. Why is that? What determines whether a democratic system has more or less effective *electoral accountability*? Just how effective can electoral accountability plausibly be? Can elections produce policies far from what the average voter would want even when voters can condition their votes on observed, actual policies? If so, why?

- For autocratic rule, it is easier to explain why things might get very messed up in the above sense. The ruling group is not accountable and so has weaker motivation to ensure broad provision of excellent public services or implementation of broadly favored policies, especially when these conflict with the ruling group's preferences and interests.

  But what explains the political failure that is autocracy? Why are autocracies not overthrown, or otherwise subject to transition to more democratic, accountable forms of government? Further, autocracy typically entails costs and major risks for the autocrats, including costly repression, risk of violent rebellion and death, and often economic policies that favor continued rule but harm overall growth. What prevents power-sharing deals that could make autocrats and their potential opponents all better off by reducing risks of violence and increasing the size of the total economic "pie" they could divide up?

- Enormous amounts of human suffering derive from large-scale violent conflicts, most of which involve armed groups seeking full control of government at the center of a recognized country,

> or in a region of a recognized country. These can be either civil wars or interstate wars. Why is power- or territory-sharing sometimes so problematic? What explains the onset of highly destructive wars, and what explains why they sometimes go on and on despite enormous costs?

A second possibly misleading aspect of what I have said so far is that while characterizing and better understanding these major political failures is a theme of the book, they will be addressed in the course of presenting basic game-theoretic methods. These are used to develop and work through arguments about the reasons for the political failures.

3 What are these models?

When things are messed up it is because people's actions are aggregating into a bad state of affairs. The models developed in subsequent sections are formal representations of *arguments* that trace out how the aggregation works in different contexts. In Thomas Schelling's (1978) terms, they formally represent verbal arguments that go from "micromotives" to "macrobehavior," providing an explanation for the latter. And because the micromotives usually presume goal-seeking, the explanation for the phenomenon at the same time provides a basis for normative evaluation.

To illustrate, here is an extremely simple example. Why is it that check-out lines in a busy supermarket tend to be of roughly equal length? (The "macrobehavior.") This one is easy. Grocery shoppers typically want to minimize the time they spend waiting in line ("micromotive"), and they can, with some accuracy, estimate current line lengths. If they individually choose the shortest line at any point in time, then lines will tend to be of equal length, since new entrants tend to go to what they estimate to be the shortest line, or will choose more or less randomly if they can't tell.

Is this a good outcome or a bad outcome? It minimizes ex ante average and variance of waiting time, which has a reasonable claim for being at least pretty good from both a utilitarian perspective (minimize average unhappiness) and also a fairness perspective (treat people equally). We could imagine trying to do even better by trying to elicit private information about costs for waiting, which could well differ across different people. For instance, post a small user fee that differs for different lines, so that people sort by willingness to pay for a shorter line. But such a *mechanism* runs up against fairness considerations and the utilitarian gains would likely be small anyway.

Here is a second example that comes from Schelling (1978) but as summarized in a 2007 *New Yorker* article (on fuel-economy standards for cars!):

> Back in the nineteen-seventies, an economist named Thomas Schelling, who later won the Nobel Prize, noticed something peculiar about the [National Hockey League]. At the time, players were allowed, but not required, to wear helmets, and most players chose to go helmet-less, despite the risk of severe head trauma. But when they were asked in secret ballots most players also said that the league should require them to wear helmets. The reason for this conflict, Schelling explained, was that not wearing a helmet conferred a slight advantage on the ice; crucially, it gave the player better peripheral vision, and

> it also made him look fearless. The players wanted to have their heads protected, but as individuals they couldn't afford to jeopardize their effectiveness on the ice. Making helmets compulsory eliminated the dilemma: the players could protect their heads without suffering a competitive disadvantage.[1]

The analysis gives us both an explanation for the initial macrobehavior – players choosing not to wear helmets despite all the damage – and a basis for both a normative evaluation and a proposal for how to improve matters. (By the way, this example also illustrates what I meant above when I said that very often things are messed up because coordinating the actions of large numbers of people is just hard. In this case coordination on a good outcome was made possible by an institution, the NHL, that could readily solve the collective action problem faced by the players. But that institution was already a work of coordination, something not easily arrived at in other contexts for a broad range of reasons.)

Neither of these two examples used any symbols or math to make the explanatory and normative arguments. And likewise all of the models/arguments presented in this book can be explained in ordinary language, something I will constantly try to do. Nonetheless the formal grammar and vocabulary is extremely useful. It is like a *scaffolding and set of tools for building arguments* in a way that makes it relatively clear just what is being said. If there is ambiguity about a verbal statement – which is very common – one can directly consult the formal representation to understand exactly what the argument is and what it presumes, to a great extent regardless of debates over what the author said about it.

A second value of the formalization is *transposability*. We get a great deal of insight into strategic and normative similarities and differences across a stunningly broad range of social, political, and economic contexts. This is already evident in the hockey example, which was invoked in an article about why things are messed up in an environmental policy problem but also transposes to Thomas Hobbes' justification for the modern state.

In a more subtle way it is also present in the supermarket-lines example, which turns out to bear a relationship to Downsian models of electoral competition and the problem of how to choose a good public policy. As discussed at length below, in democracies candidates for office who compete by committing to policy positions can have incentives to cater to the median voter. This can be given a normative justification as a good outcome on similar grounds to checkout lines – that it approximately minimizes the typical voter's unhappiness with policy.

As in the supermarket example, in principle we might do even better if we could elicit private information that people have about how intensely they care about different policies. And any such scheme faces a similar obstacle: Because of incentives to misrepresent intensity of preference, costs are going to be necessary to get "truthful revelation" (like the hypothetical user fees for different supermarket lines). Lobbying, voter mobilization drives, campaign contributions, petitions, protests and marches, organizing social movements, armed rebellion in some cases – all of these basic features of politics can be understood as costly responses to the inherent limitations of the electoral mechanism for revealing citizen preferences. And just as with supermarket lines or road tolls, there are tradeoffs and tensions between fairness considerations and maximizing average welfare. Fairness: Why should having more

[1] James Surowiecki, "Fuel for Thought," *The New Yorker* 23 July 2007. By the way, Schelling's analysis of the hockey helmet problem is exactly the same as Thomas Hobbes' core argument for why government as a third-party "Leviathan" is a good thing.

money, or more inclination to go demonstrate in the streets, give you more political influence? Average welfare: Why shouldn't those who care intensely about a particular policy be able to contribute time and money to influence the outcome?

A common mistake is to assume that if one uses a formal model to make an argument about some phenomenon in the human world, the modeller must believe that there is no essential difference between the natural and human worlds, and that social science can and should be "like physics."

It is true that formal models in the natural sciences are usually explanatory in a similar way: They are also reductionist. They posit some entities that "behave" according to some rules, and then explain by showing how higher-level, macro phenomena follow from the aggregation and interaction of the micro-level "behaviors" (for instance, water boiling, rain, fevers, chemical and nuclear reactions, . . .).

But beyond this similarity there are fundamental differences. Models in physics and chemistry purport to describe time-invariant laws of nature that should hold anywhere and everywhere. They refer to entities like atoms that are not reflective, not intentional (goal-seeking), and do not try to anticipate and condition on the behavior of other like entities. Because the entities are not goal-seeking the models have no normative content. There is no sense in scolding an atom for how it acts, or for assessing how well some physical interaction achieves the atoms' ends (they don't have ends). The entities in natural science models cannot reflect on the explanation for a bad state of affairs and deliberately seek to rearrange incentives or goals to make things better.

Because atoms (etc.) do not anticipate and condition how they behave on expectations about others' actions, the phenomenon of *multiple stable states* (aka multiple equilibria) is far less evident in physics and chemistry than in the human world. This is actually an implication or result of game-theoretic analysis; see below. A related implication is that there are classes of situations in which human "macrobehavior" will be unpredictable even when the "micromotives" are highly regular – and not because of complexity, which is standard in both natural and social domains, but because of anticipatory strategy.

4 Theory and empirics: Explanatory versus black-box models

How does one assess whether the explanation given by the argument that a model makes is a good one, empirically? There is no short answer, because there are many different sorts of models/arguments developed for many different contexts with enormous variation in types of evidence.

Most of the models considered in this book are essentially *thought experiments* that ask about what would follow at the level of some macrobehavior if we assume x, y, z, about micromotives and strategic context. As such they can be clearly wrong but still be invaluable as empirical tests!

We observe some pattern of actions in the world, and then hypothesize that it might be explained as the result of such-and-such micromotives aggregating in such-and-such a fashion. We then work through the model/argument to see if this is so. It may not be, in which case we need to go back to the drawing board. But we do so with a better understanding of what theory of the "data-generating process" (DGP) can and cannot explain the empirical observations (or in other words, what is a feasible explanation).

Often the model/argument will get some things right about the macro-level empirical patterns, but other things wrong or unclear, which naturally provokes further theoretical development or empirical investigation. And of course even a relatively good match to empirical patterns does not guarantee that this is the only or the most important explanation for how they arise. So empirical assessment of assumptions of the model/argument, testing implications besides the main implications, and proposals for alternative explanations may be warranted.

It is a back-and-forth process. Sometimes it leads all the way from simple thought-experiment models to "structural models" of a data-generating process whose parameters can be estimated empirically. This is common in Economics but remains rare in Political Science.[2]

The idea of social science research proceeding as a back-and-forth between models of DGPs and empirical assessments of their assumptions and implications should be a commonplace for political scientists. These days it is not, or not so much. Instead, a common view widely taught in methods courses is that one is doing political science if and only if one is constructing and executing research designs that "identify causal effects."

To illustrate just how different the accumulate-causal-effect-estimates conception is, note that the back-and-forth approach can advance, in principle, simply by making sense of observed correlations in the world, none of which are causal. This is entirely standard in natural science. For instance, Newton, Darwin, and Einstein each developed models of DGPs that made better sense of observed correlations than did prior models. None of them estimated causal effects after the current fashion, or used causal effect estimates as critical inputs to their DGP models.

Causal effect estimates can be useful, and, in a weak sense, identifying a cause is a form of explanation. Why did his headache go away? Probably because he took aspirin and aspirin causes a reduction in headaches on average. But while an experiment that estimates the causal effect of aspirin on headache reduction can be valuable, it can tell us literally nothing about *why* aspirin causes headache reduction. The actual science would lie in theoretical and empirical work that elucidates the micromotives-to-macrobehavior biochemical processes that go from aspirin in the bloodstream to perceived pain reduction. Without the latter, we would be reduced to randomly grabbing plants to try for treating diseases, and we would have no basis for expectations about whether or how causal effect estimates from one experiment would manifest "the next time" (in any other setting).[3]

To develop this point further using a social science example, here is a parable.

Imagine a world of social scientists who have *zero* conception of how prices are determined in goods markets. They have not developed the theoretical ideas of supply and demand, or the idea that prices and quantities are mutually determined "in equilibrium."

These researchers want to understand the causal effect of prices on the amounts of a good that is sold, and they are well-versed in methods of causal inference. They have already tried collecting data on prices and quantities sold in different markets and in different periods of time. For instance, perhaps

[2]For an excellent and instructive example on a political economy topic, see Francois, Rainer and Trebbi (2015).

[3]In many cases, the results of drug trials on one set of human subjects can be expected to be externally valid for any other set of subjects, because human biochemistry is known to be similar from person to person. This is emphatically not the case for almost any political science question you can think of.

they have observations (p_t, q_t) for prices of gasoline, and amounts of gasoline sold, for different time periods $t = 1, 2, \ldots, n$. Or instead the t's could index different locations.

The researchers have already noticed that there is no appreciable or consistent correlation between p_t and q_t. Regressing q_t on p_t yields an approximately zero coefficient. But they have taken methods courses, and they know both that correlation does not imply causation and that lack of correlation does not imply absence of a causal effect. There could be confounds, also known as omitted variables, that are obscuring the true "average treatment effect" of prices on quantities sold.

So they undertake to look for natural experiments, here meaning situations where there is a strong argument that either prices or a major influence on prices (that does not also directly affect quantities) varied essentially at random.

Unfortunately, the results of these efforts are not consistent across studies. In some – for example, in several that argued that international oil prices vary essentially at random, and used these as "instruments" for gas prices in a location – the studies tended to find a negative effect of prices on quantities sold. In others – for example, in several that argued that prices tended to fluctuate randomly with seasons, or weather – the studies tended to find a positive effect of prices on quantities sold. In yet others, the null hypothesis of zero causal effect could not be rejected.

A review essay concluded that despite initial excitement generated by a "first well-identified study" that had found a negative effect, this did not hold up when researchers asked about "external validity," examining other contexts with different identification strategies. Surprisingly, there did not seem to be any consistent causal impact of prices on aggregate quantities sold. Given this, it was judged unlikely that an increase in gas prices would reliably lead to less gas consumption and thus lower carbon emissions.

The review noted that it was puzzling that some studies at the level of individual gas stations had found that randomly increasing the posted price correlated with lower sales by that gas station. But it was thought that this indicated that great care was needed in trying generalize from one level of analysis to another. One could not generalize from the micro-level to aggregate effects, probably because of complex, non-linear interactions, network effects, or maybe "emergent phenomena."

It was widely agreed that the problem could be solved, or, at least, that great progress could be made, by applying the gold standard: a randomized controlled trial. Some researchers managed the politically and logistically challenging feat of convincing several local governments to legislate that gasoline prices in their areas had to be set at levels that would be randomly chosen by the researchers. All gas stations in the area would, by law, have to post and charge these randomly set prices for the period of time of the study.

There were huge problems in implementation, mainly concerning compliance. Some doubted whether the results were meaningful for this reason. But there were results: Remarkably, it appeared that if anything, the average treatment effect of prices on quantities was *non-linear* – inverse U-shaped, in particular. Quantities sold appeared to increase up to a point, and then decline. It was speculated that past results may have reflected studies that "identified" on one or the other side of the relationship, or got null results by being in the middle.

What's going on here? If you've taken Econ 101 and been introduced to the theoretical concepts of supply and demand curves – which were hard-won theoretical innovations over a several centuries – you are likely to see the whole program of these empiricists as confused and unproductive. Having or developing an explanatory, micromotives-to-macrobehavior model of the DGP for prices and quantities would have been very helpful.

Supply and demand are theoretical concepts – you can't see them in the world anywhere. Supply is the hypothetical amount of goods that would be offered for sale at given price, and demand is the hypothetical amount of goods that people would seek to buy at a given price. From more micro-models of consumer and firm decision-making, we get the implication that, normally – based on assumptions about consumer preferences and firm production technologies – we should expect supply curves to slope up and demand curves to slope down. Different micro models can be fleshed out concerning the specifics of how adjustment of prices and quantities works, but the implication will be that prices adjust so that demand equals supply (else prices are bid up and demand falls while supply rises, or vice versa).

Summarizing formally, the DGP model has two equations, one for the supply and one for the demand curve in a given market or time period t. Illustrating with a simple linear version:

$$\begin{aligned} q_t^D &= a_t - bp_t \\ q_t^S &= c_t + dp_t, \end{aligned}$$

where S = supply, D = demand. a_t and c_t are other influences on amounts demanded and supplied (besides price) at time t. $b > 0$ and $d > 0$ are *structural parameters*, meaning that they are assumed to be fixed across t's.[4] Notice that with $b > 0$, the quantity demanded decreases with price, while $d > 0$ would imply that the quantity offered to sell increases with price. In a market equilibrium, $q_t^S = q_t^D = q_t^*$. We can then solve for equilibrium levels of price and quantity in t as

$$\begin{aligned} q_t^* &= \frac{1}{d+b}\left(a_t d + c_t b\right) \\ p_t^* &= \frac{1}{d+b}(a_t - c_t). \end{aligned}$$

Observe that p_t^* and q_t^* will be positively correlated through the a_t's, which are the unmeasured exogenous influences on *demand* for the good, but negatively correlated through the unmeasured c_t's, which are the unmeasured exogenous influences on *supply* of the good. A positive "shock" to demand causes more people to try buy at the current price, leading to price increase and also an increase quantities sold. A negative shock to supply (from, for instance, supply disruptions like bad weather causing poor harvests) reduces amounts at market, bidding up prices, which partially but not completely mitigates the reduced supply effect.

This model/argument makes sense of the empiricist's findings.

- The simple correlation between p_t and q_t is indeterminate. It can go either way, but perhaps tends towards zero since the two types of exogenous determinants are offsetting.
- Instruments for p_t that capture some part of random variation in demand (that is, measures of a_t)

[4] Or we are hoping to estimate an average value.

will "identify" movement along the supply curve, thus showing a positive relationship. Instruments for p_t that capture some part of random variation in influences on supply (measures of c_t) will identify based on movement along the demand curve, thus showing a negative relationship. The relative size of these effects will depend on the structural parameters b and d, which reflect more micro-level facts about consumer preferences and firm or industry costs and production technologies.[5]

- If one tries to force prices to take a certain level (in the randomized controlled trial), then either demand will exceed supply at that level, or supply will exceed demand (except close to the equilibrium levels). In the first case, many suppliers will not cover their costs adequately and so total sales will be less than at the equilibrium level.[6] In the second case, the consumers simply won't buy at that price as much as they would at the equilibrium level. Thus, an inverse U-shaped relationship would appear in the RCT data.

From the perspective of the supply-demand model, the question that the empiricists started with – "What is the causal effect of prices on quantities sold?" – no longer makes sense, or at least is badly posed. In the explanatory model, prices and quantities are mutually determined as a result of a micromotives-to-macrobehavior process. So it doesn't really make sense, or at a minimum is confusing, to talk about the causal effect of prices on quantities. The "what is the average causal effect?" question is implicitly assuming that there is some sort of independent or manipulable kind of variation in prices. The supply-demand model rejects that premise concerning the determination of prices in a competitive market, while at the same time predicting what would be the causal effect of attempts to legislate prices.

Put differently, the parable illustrates that the "what is the causal effect of X on Y?" or "what is the average treatment effect of X?" questions are premised on a *black-box model* of the data-generating process. There is an unspecified process, a black box, interposed between the X and the Y. Getting an estimate of the average treatment effect tells us nothing by itself about what is in the black box, that is, what explains why and how there is a relationship. If and when we have an *explanatory model* of the data-generating process, we are going to be interested in estimating its structural parameters rather than average causal effects.

Further, when a political scientist wants to estimate the causal effect of some X on some Y, the X is usually itself something that is endogenously determined in a political economy or social equilibrium, like prices and quantities – for example, a policy or law, a constitutional feature like an electoral system, a campaign strategy, the race of a survey respondent, a survey respondent's opinion about an issue, democracy, a signature on an international agreement, and so on. As the supply-demand example shows, in an equilibrium the effect of variation in one exogenous influence on an endogenous X can be very different from the effect of variation in another exogenous influence. Without a model of the DGP, we have no basis for claiming that a "local average treatment effect" generalizes, or even interpreting what we are finding.[7]

[5] In Economics, the term "identification problem" originally referred to the problem of how to estimate structural parameters of a DGP model (like b and d) from data. Note how this differs from the current term "causal identification," which refers to a focus on, and methods for, distinguishing correlation from causation.

[6] The model additionally predicts rationing and compliance problems ("corruption") due to excess demand at prices that are legislated to be lower than equilibrium level.

[7] To be painfully clear: I am saying that most political science is in the position of the hypothetical social scientists in

Evidence about causal effects can still be very useful for developing or evaluating explanatory models of DGPs. In particular, an explanatory model may make a prediction about a causal effect that can be assessed with an experimental or quasi-experimental research design. Causal effect estimates can be valuable inputs into the back-and-forth process of developing explanatory DGP models (Nakamura and Steinsson, 2018). But this is also true for mere correlations. For instance, we ask if measurable correlations are as implied by the DGP model. Causal effect estimates in addition usually stimulate "why?" questions and efforts to develop explanatory models, which will invariably involve using information about correlations other than the "well-identified" one (e.g. Fearon, Humphreys and Weinstein, 2015).

5 What is the point of political science?

The intent of this short section is context. Where and how does the material that follows fit within the larger projects of political science? What are those larger projects? Why are we doing this?

We can have many good reasons for studying politics, including the same kind of curiosity that motivates disciplines such as astrophysics. Nonetheless, I would put front and center

the constructive evaluation of the performance of political systems.

To evaluate performance we need two things. First, normative standards. What constitutes good or bad performance? Second, we need systematic work to measure performance by these standards.

To be constructive, we need to be able to *explain why* a political system is performing well or poorly. Worthwhile proposals for, or attempts at, reform and improvement must be premised on a theory about why the current system is producing bad outcomes in some respect – that is, why things are messed up. Political science research is more constructive the more it provides usably-accurate explanations for why a political system produces the good or bad outcomes that it does.

In sum:

1. Normative standards for government performance.
2. Empirical work to measure performance by these standards.
3. Coherent, assessable explanations for why the measured performance is good or bad in different areas, in the form of models of the data (or performance)-generating processes at issue.

This book develops a set of ideas and arguments about (1), applying these while working through and developing "workhorse models" relevant to (3). I comment here and there on (2), mainly in regard to the performance of the US political system. But I do not engage in a systematic attempt to measure its performance, or that of any other political system, by the normative standards advanced in Part III.

the parable. I am not saying that there are social scientists who are or were doing research on prices and quantities like that described in the parable.

[. . . contrast to post-45 "behavioral" and "positive" Political Science, which does not try to systematically measure performance in part because it has been so chary of explicit discussion, conceptualization, or use of normative standards for evaluating performance. . . .]

Part II

Game theory starter

6 Strategic situations

The term game theory refers to a large set of concepts, arguments, and formal representations developed since the 1950s, all pertaining to the analysis of *strategic situations*. Strategic situations are found in an enormous variety of specific contexts, both human and non-human, so it is not surprising that game-theoretic ideas and analyses are now important in many fields of inquiry. These include Political Science, Economics, Psychology, Biology, and Computer Science.

What is a strategic situation? Any situation that can be represented as having

1. A set of *agents* who take actions. These might be individual people; organizations that have a way of making a choice; animals that are "programmed" by genetics to act in a certain way in a certain context; or computer programs, like those governing autonomous vehicles. We will usually denote the set of agents as a set $I = \{1, 2, 3, \ldots, i, \ldots, n\}$, using numbers for the agents' "names," and using i for a typical or arbitrary agent.

2. A set of *actions* available to each agent, represented (in this book, usually) as $A_i = \{a_i^1, a_i^2, \ldots, a_i^{m_i}\}$. Here, the superscripts are distinguishing different actions available to agent i, who has a total of m_i actions.

   For example, for drivers arriving at an intersection with four-way stop signs, a simple depiction of available actions could be $A_i = \{Go, Wait\}$ for $i = 1, 2$.

   Or consider a majority-rule election between two candidates, labelled (say) D and R. An individual voter's strategy set could be $A_i = \{D, R\}$, or $A_i = \{D, R,$ don't vote$\}$ if we allow for a turnout decision.

   We will constantly use $a_i \in A_i$ to mean some particular action in A_i.

   The actions or strategies $a_i \in A_i$ – I will use "action" and "strategy" interchangeably most of the time – are called *pure strategies*. Later we will often use the idea of a *mixed strategy*. A mixed strategy is a probability distribution on the set of pure strategies. For example, flipping a coin and using the result to decide whether to choose a_i^1 or a_i^2 would be a literal mixed strategy. We denote the set of all possible probability distributions on a (finite) set of actions A_i as $\Delta(A_i)$, with typical element (mixed strategy) $\alpha_i \in \Delta(A_i)$, which is a list of probability numbers that are the probability with which each corresponding pure strategy is played. More on this later, don't worry if it is confusing or odd now.

3. A set of *outcomes*, which are states of affairs in the world that are the result of the agents taking a given combination of actions. The set of all action combinations is represented by the Cartesian product[8] of the agents' action sets, thus

$$A = \times_i A_i = A_1 \times A_2 \times \ldots A_i \times \ldots A_n.$$

In the driving example, $A = A_1 \times A_2 = \{(Go, Go), (Go, Wait), (Wait, Go), (Wait, Wait)\}$.

Notice that we are associating *outcomes* with action combinations – e.g., (Go, Go) might associate with a car crash, or the outcome that has a possibility of an accident.

In the election example, the set A is the set of all vectors (lists) like

$$a = (D, R, D, R, R, \ldots, D)$$

where each entry is what that voter chose. Notice that this set has 2^n elements, which is huge for a large electorate. Despite the large number of action combinations, in terms of outcomes we care about, there are just two sets (if n is odd): those where more people voted for D, and those where more people voted for R. But the elements of the set A map uniquely into the outcomes over which, under typical assumptions, the voters have preferences.

4. For each agent, a set of *preferences* over the set of outcomes A, meaning judgements of the form "I prefer this outcome to that one," or "I like this outcome at least as much or better than that one." For notation, if a and b are two outcomes in A – that is, $a, b \in A$ – we write

$$a \succeq_i b$$

to mean that agent i *weakly prefers* outcome a to outcome b. In other words, i likes outcome a at least as much as outcome b, or strictly better. We will use

$$a \succ_i b$$

to mean that i *strictly prefers* outcome a to b, and

$$a \sim_i b$$

to mean that i is *indifferent*, or doesn't care between a and b.

Notice that $\succeq_i$ is **not** a $\geq$ sign: a and b are outcomes in the world, not numbers! It is just a shorthand way of writing "i (weakly) prefers ...."

Although the meanings and relationships between strict and weak preference and indifference are intuitively clear, they are formally defined in terms of each other as follows.

Definition 1. *For $a, b \in A$, $a \succeq_i b$ if and only if it is not the case that $b \succ_i a$; and $a \sim_i b$ if and only if $a \succeq_i b$ and $b \succeq_i a$.*

[8] Recall that the Cartesian product of two sets is the set of all possible pairs where the first element in the pair taken from the first set in the Cartesian product and the second element from the second set. For example, if $M = \{turkey, ham\}$ and $B = \{wheat, rye\}$, then $B \times M = \{(wheat, turkey), (wheat, ham), (rye, turkey), (rye, ham)\}$. Note that $M \times B$ differs in that the first element of each pair is a meat, and the second a bread type.

For the drivers-at-an-intersection example, plausible or typical preferences over outcomes are pretty obvious. For the election example, a typical approach would be to suppose that each voter i has a preference for either D or R, so that $a \succeq_i b$, $a, b \in A$, just when total votes for i's preferred candidate are greater in a versus b.

We will almost always suppose, further, that each agent can order the outcomes from most preferred (best) to least preferred (worst). More on this assumption below.

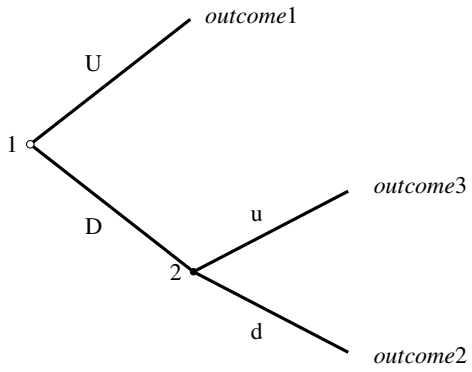
Digression: Humans, organizations, and autonomous vehicles can (often) be understood to have preferences over outcomes, but what about non-human animals or, say, bacteria or viruses? In those cases, instead of preferences over outcomes, game-theoretic analyses will often proceed by having the *reproduction rate* of an agent depend on the outcome $a \in A$ that occurs. Then an agent's genetically programmed action does better, in terms of its frequency in the next generation of animals/agents, the higher its reproduction rate, and reproduction rates depend on outcomes (other animals' "choices", as well as one's own). It turns out that in this approach, common in evolutionary game theory, reproduction rates play a role similar to the preferences of agents whom we think of as having intentions.

Formalism note: Although you don't need think about it this way, formally a person's preferencs over outcomes $\succeq_i$ can be thought of as a *set*, namely, $\succeq_i \subset A \times A$. That is, i's preferences are the subset of $A \times A$ such that it is true that for each $(a, b) \in \succeq_i$, the first element is weakly preferred to the second element. For example, later we will say that a person's preferences are *complete* if for all $a, b \in A$, either $a \succeq_i b$ or $b \succeq_i a$. Equivalently, for all $a, b \in A$, $(a, b) \in \succeq_i$ or $(b, a) \in \succeq_i$.

Those four elements – agents (or players), actions for each agent, outcomes that result from action combinations, and preferences over outcomes for each player – are the raw ingredients of a strategic situation.

What about time and sequence? In the super-abstract framework as developed so far, we are in effect imagining the players choosing their actions *simultaneously* or *in ignorance of* each others' choices. But you know well that many, or indeed close to all, situations of human interaction in the real world have an element of sequence and time. You choose an action now, in anticipation of what someone else might do next, after seeing your action, for example.

Later on we will analyze another way of representing strategic situations, using tree diagrams like this:



This is called an *extensive-form* game, whereas the framework developed so far is called a *normal-* or *strategic-form game*. You can see that an extensive-form has the same basic elements: players, who have sets of action choices, which combine to produce outcomes, over which the players may have preferences. What is different is that an extensive form encodes additional information about the temporal order of actions, and also about who knows what concerning what has happened at each point in game time.

For a number of reasons it is best to begin by studying normal-form games, and then later move to new issues that arise when we introduce time and dynamics. This is what we will do.

7 Solution concepts, Nash equilibrium

As described so far, this definition of a generic strategic situation can be summarized formally as a triple

$$\Gamma = \langle I, (A_i), (\succeq_i) \rangle,$$

where we are using the parens to indicate a list of components, one for each $i \in I$. This is a definition of a normal-form game: Players, strategies, and preferences for each player over all combinations of strategies (outcomes).

Clearly, this is an extremely abstract, context-free formulation. By moving to this high level of abstraction, game theory has allowed insights into and understanding of fundamental similarities that appear across highly diverse contexts that no one had previously thought to connect. For example, basic similarities between situations involving bacteria, or spiders, and people or organizations like states, or among internet servers. Or, just within the human domain, between situations in politics, economics, or daily social interactions that were previously understood as entirely unrelated. This can be a plus.

But we are going to have to start to add contextual specifics back to get value for understanding strategic situations in particular domains. In this book we focus on game models developed for better understanding politics. Nonetheless, it is useful to begin at the high-level of abstraction, by way of introduction.

Given a normal-form game, natural questions to ask are, How *should* and how *will* a player choose among actions, assuming (to start out) knowledge of Γ, which includes knowledge of other players' preferences and the mapping from action combinations to outcomes?

Note that I pose both the normative question (should) and the positive question (will). The normative question asks about how a player who wants to do best in light of their preferences $\succeq_i$ should choose, given this end. We will typically refer to this under the heading of what is it "rational" for a player to do.

The positive question, how will a player choose, admits a broader variety of interpretations, including ideas about how boundedly rational or psychologically-biased players (relative to the normative standard of what would be "rational") would in fact choose. Notice that we need to understand the normative standard of what it would be rational to choose first, in order to be able to identify what

psychological bias, or anything not "rational," would be.

At the core of game theoretic methods are a series of proposals about how to answer the normative and positive questions, which go under the general heading of *solution concepts*. Loosely put, a normal-form solution concept is a proposal for how to pick out a subset of action combinations $a \in A$ that have some appealing properties – as formalized by the solution concept – perhaps as predictions about what would or should happen for an arbitrary normal-form game Γ.

"Appealing properties"? What might these be? Overwhelmingly in what follows and in applied game theory work in general, the core "appealing property" is some notion of *stability*. Stability in the sense that if players expect that action profile $a^* \in A$ is going to occur, each player i still wants to choose the action a_i^* that is a component of a^*.

This kind of stability, or *Nash equilibrium*, is a reasonable expectation to have or prediction to make for any *recurrent pattern of behavior*. If a pattern of behavior (that is, an "outcome") occurs over and over – everyday, for instance – then people have common expectations about what's going to happen and must have adjusted their own actions to the anticipation of this pattern. This makes Nash-equilibrium-type solution concepts natural and useful for analysis of repeated patterns of outcomes in particular domains, such as, for instance, car traffic or internet-routing patterns, market behavior by consumers, legislative politics, or various aspects of electoral politics.

But as we will discuss, the workhorse solution concept of Nash equilibrium does not, by itself, supply any story or theory for why or how players of a particular game Γ would coordinate their expectations on any one particular Nash equilibrium – there can be more than one, as we will see below. Relatedly, the assumptions of individual rationality and common knowledge of the strategic situation turn out not to be sufficient to guarantee that rational players would necessarily coordinate on Nash equilibria. More on this point later.

Back now to the question, How should you choose among your available actions A_i if you are player i in the situation represented as the normal-form game $\Gamma = \langle I, (A_i), (\succeq_i) \rangle$? This can be extremely difficult, it turns out, largely because the other players, like you, are trying to answer the same question, so that everyone's best choice may depend on their expectations, or guesses, about everyone else's. So there can be a confusing recursion to deal with.

Let's start with some basic concepts and definitions.

Let $a_{-i} = (a_1, a_2, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n)$ be the vector (or list) of actions chosen by *everyone besides player i*. So we can write $a = (a_i, a_{-i}) \in A$ to indicate a complete action profile, but breaking out player i's action a_i from the rest.

If (somehow!) player i expected that the other players were going to choose the list of actions a_{-i}, then normatively it seems that if i wanted to do the best she could do in light of her preferences $\succeq_i$, she should choose an action that yields her most preferred outcome given a_{-i}.

Definition 2. *a_i is a* **best reply** *to a_{-i} if*

$$a = (a_i, a_{-i}) \succeq_i (a_i', a_{-i}) \text{ for all } a_i' \in A_i.$$

We can also define a **best reply correspondence**[9] *for i as the set*

$$BR_i(a_{-i}) = \{a_i \in A_i : a_i \text{ is a best reply to } a_{-i}\}.$$

Of course, we have no theory or idea yet about how i should or would form beliefs about what the other players are going to choose. But even lacking this, we can still draw some weak conclusions about what a rational player should *not* do. If an action a_i is not a best reply for *any* a_{-i}, then no matter what i thinks the other players are going to do, it seems reasonable to say that a rational player i should not want to choose a_i. Thus,

Definition 3. a_i *is* **never a best reply** *if* a_i *is not a best reply for any* $a_{-i} \in A_{-i}$.

"Never a best reply" is one notion of what a bad action choice would be. Here is another: A rational player should not want to choose an action a_i if there is another action available that gives a strictly more preferred outcome *no matter what* the other players choose.

Definition 4. a_i' **strictly dominates** a_i *if* $(a_i', a_{-i}) \succ_i (a_i, a_{-i})$ *for all* $a_{-i} \in A_{-i}$. *We also say, in this case, that* a_i *is* **strictly dominated**.

[*Technical note that you can skip if you wish.* Clearly, if an action a_i is strictly dominated by some other action a_i', then a_i is never a best reply. But a strategy a_i can be never a best reply even if there is no other pure strategy that strictly dominates it.[10] In two-player games, a strategy is never a best reply if and only if there exists some other pure *or mixed* strategy that strictly dominates it. In games with more than two players, this is still true but we have to allow not just for independent mixed strategies, but for correlated mixed strategies as well.]

So these are two quite reasonable proposals for characterizing actions that a rational player would *not* want to choose, for our extremely general, abstract characterization of strategic situations. And we will see that "never a best reply" and "strictly dominated" can in fact generate sharp predictions and interesting implications for certain strategic situations – situations that have an aspect of a *commitment problem*, to be described below. For a great many other strategic situations, however, "never a best reply" and "strictly dominated" will not get us very far at all. This is the case for the large and interesting class of strategic situations that have an aspect of a *coordination problem*, to be defined below.

Armed with the concept of a best reply, we can now give our first version of the definition of the workhorse "stability" solution concept, Nash equilibrium. In words, an action profile $a^* \in A$ is a pure-strategy Nash equilibrium if each player is choosing a best reply to what every other player is choosing.

[9] A correspondence is like a function, but whereas a function $f(x)$ has a unique value for any input x, a correspondence can have multiple values. It is a "set-valued function." We need it here because there can be more than one best reply, since a person might be indifferent between two or more actions that are equally good and better than any other option.

[10] For example, in this game, R is never a best reply to U or D, but it is not strictly dominated by L or C. (Think of the numbers as dollar payoffs for now, with the first one being the row player's payoff, and the second being the column player's payoff.)

	L	C	R
U	10, 0	0, 10	3, 3
D	0, 10	10, 0	3, 3

Definition 5. *$a^* \in A$ is a* **pure-strategy Nash equilibrium** *if for all i, a_i^* is a best reply to a_{-i}^*. That is, for all i, $a^* = (a_i^*, a_{-i}^*) \succeq_i (a_i, a_{-i}^*)$ for all $a_i \in A_i$.*

Equivalently,

- *$a_i^* \in BR_i(a_{-i}^*)$ for all i; or*
- *defining $BR : A \to A$ as the list of best-reply correspondences for each player,*[11] *a pure-strategy Nash equilibrium is an action profile $a^* \in A$ such that $a^* = BR(a^*)$. In words, each player is choosing an action that is a best reply to what all the other players are choosing.*

The idea of Nash equilibrium is simple and straightforward. But in my experience it can take students a long time to fully appreciate and apply it to artificial problems and real-world situations. We will start with some simple examples, after stating a weak criterion for assessing outcomes in a strategic situation as good or bad, better or worse.

8 Good and bad outcomes, a first cut

From the perspective of an individual player, it is easy to say what constitutes a good or bad outcome in a strategic situation – outcome *a* is better for *i* the higher *a* is in *i*'s preference ordering. But what can we say, if anything, about what constitutes a good or bad outcome for *all* of the players, in some sense of collective welfare or wellbeing? If we want to explain why things in politics are so messed up and what might be done to make them less messed up, we have to articulate criteria for evaluating whether an outcome in a strategic situation is good or bad in some overall sense.

This is a hard problem that necessarily gets us into moral philosophy if we go at it properly.[12] Although this is definitely not a book on ethics, we will return to the issue at some length when we start talking about evaluating what is a good or bad policy outcome (Part III). For now, here is an important and useful ethical criterion that is defensible in a great many, though definitely not all, situations.

Definition 6. *Outcome $a \in A$* **Pareto dominates** *outcome $b \in A$ if for all $i \in I$, $a \succeq_i b$ and for some i, $a \succ_i b$. An outcome is* **Pareto dominated** *if there is another outcome that Pareto dominates it. An outcome is said to be* **Pareto efficient** *if it is not Pareto dominated, and* **Pareto inefficient** *if it is Pareto dominated. We will often just say "efficient" or "inefficient."*

In words, the Pareto criterion says that one outcome is better than another if all agents like it at least as much, and at least one agent likes it strictly better. In many settings – in fact, in pretty much any setting where there is good reason to see the agents' preferences as worth respecting – this seems like an unobjectionable standard.

The Pareto criterion spotlights the moral intuition that it is not good to "leave money [in general, wellbeing] on the table." But it has nothing to say relevant to common moral intuitions about *fair*

[11] That is, $BR(a) = (BR_1(a_{-1}), BR_2(a_{-2}), \ldots, BR_n(a_{-n}))$.

[12] Indeed, even at the individual level, no one believes that the normatively best outcome for a person in any given strategic situation is necessarily the outcome that they most prefer.

distribution. If you and I have \$100 to divide, and we both prefer more money to less, then *any* division of this "pie" that sums to \$100 is Pareto efficient, including the division that gives me all of it and you nothing (and vice versa). $(60, 40)$ Pareto dominates (e.g.) $(55, 30)$. But $(60, 40)$ and, say, $(50, 50)$ are not **Pareto ranked**, even if there are many situations in which people think $(50, 50)$ is a socially better outcome than $(60, 40)$, or $(100, 0)$.

9 Taxonomy of strategic situations

By developing a general, abstract representation of "strategic situations," game theory made possible a general taxonomy of types of strategic situations. As noted above, this has allowed us to see interesting and, arguably, fundamental similarities and differences across a stunning variety of contexts. It has also greatly facilitated the analysis of reasons for bad outcomes in many contexts, along with what is needed to make things better.

For strategic situations modeled as having two agents with finite strategy sets A_1 and A_2, we can conveniently represent a game using a matrix like this:

		2	
		L	R
1	U	10, 10	0, 15
	D	15, 0	3, 3

Here player 1 is choosing between two actions labelled U and D, and player 2 is choosing between two actions labelled L and R. Thus $A_1 = \{U, D\}$ and $A_2 = \{L, R\}$. The four cells in the matrix correspond to $A = \{U, D\} \times \{L, R\} = \{(U, L), (U, R), (D, L), (D, R)\}$. The first number in each cell is a "payoff" for player 1, the second a payoff for player 2.

9.1 Representing preferences

We will have more to say about the meaning of these numbers later on, but for now (and in fact almost all of what you need to know): These are just a convenience for us, the analysts, to represent the actual primitives, which are the preferences over outcomes, $\succeq_1$ and $\succeq_2$. Suppose that the agents in our abstract description of a strategic situation are able to *order* the outcomes from their most preferred to their least preferred. Here, we are just using the numbers to conveniently represent the ordering given by i's preferences, $\succeq_i$.

Formally, the numbers for player i are a *utility function* $u_i \colon A \to \mathbb{R}$, which is to say, a function that assigns a number to each outcome in A. $u_i(a)$ is the number assigned by this utility function to the outcome $a \in A$.

Definition 7. *A utility function $u_i \colon A \to \mathbb{R}$* **numerically represents** *preferences $\succeq_i$ if for any two outcomes $a, b \in A$, $a \succeq_i b$ if and only if $u_i(a) \geq u_i(b)$.*

For instance, in the above model, $u_1(U, L) = 10$, which is greater than $u_1(D, R) = 3$, which says nothing more than that player 1 strictly prefers the outcome (U, L) to (D, R), or $(U, L) \succ_1 (D, R)$.

Observe that this implies that *the specific number assigned to any given outcome has no meaning whatsoever by itself.* We could assign any of, say, 6, 0, or -1,000,000 to $u_i(U, L)$ and it wouldn't matter which, as long as we changed the other values of $u_i(a)$ so that the ordering was preserved.

What has to be true about a person's preferences for there to be a numerical representation of them, in this sense?

Definition 8. *A person's preferences $\succeq_i$ over a set of outcomes A are* **complete** *if for all $a, b \in A$, either $a \succeq_i b$ or $b \succeq_i a$. A person's preferences $\succeq_i$ are* **transitive** *if, for $a, b, c \in A$, $a \succeq_i b$ and $b \succeq_i c$ implies $a \succeq_i c$.*

Theorem 1. *For a finite set of outcomes A, there exists an ordinal utility function that numerically represents $\succeq_i$ on A if and only if $\succeq_i$ are complete and transitive.*

Proof. Only if: Since the numbers $u_i(a)$, $a \in A$, are complete and transitive, then u_i represents $\succeq_i$ implies that $\succeq_i$ is complete and transitive. If: Let $u_i(a)$ equal the number of outcomes $a' \in A$ such that $a \succeq_i a'$. By completeness, $u_i(a)$ exists for all a. By transitivity, if $a \succeq_i a'$, then $u_i(a) \geq u_i(a')$, since for all a'' such that $a' \succeq_i a''$, it is also true that $a \succeq_i a''$. Last, suppose to the contrary that $u_i(a) \geq u_i(a')$ and $a' \succ_i a$. This makes for a contradiction. It must be that $u_i(a') > u_i(a)$ because by transitivity, $a' \succ_i a \succeq_i a''$ for all a'' such that $a \succeq_i a''$, and $a' \succ_i a$ adds one more to $u_i(a')$.

Digression on "rational": Preferences $\succeq$ on a set of actions or outcomes A that are complete and transitive are often termed "rational." I would prefer to associate "rational" with the idea of choice rather than conditions on preferences. In terms of the concepts we have introduced so far, a person is rational if, when they choose $a \in A$, there is no $b \in A$ such that the person strictly prefers b to a.

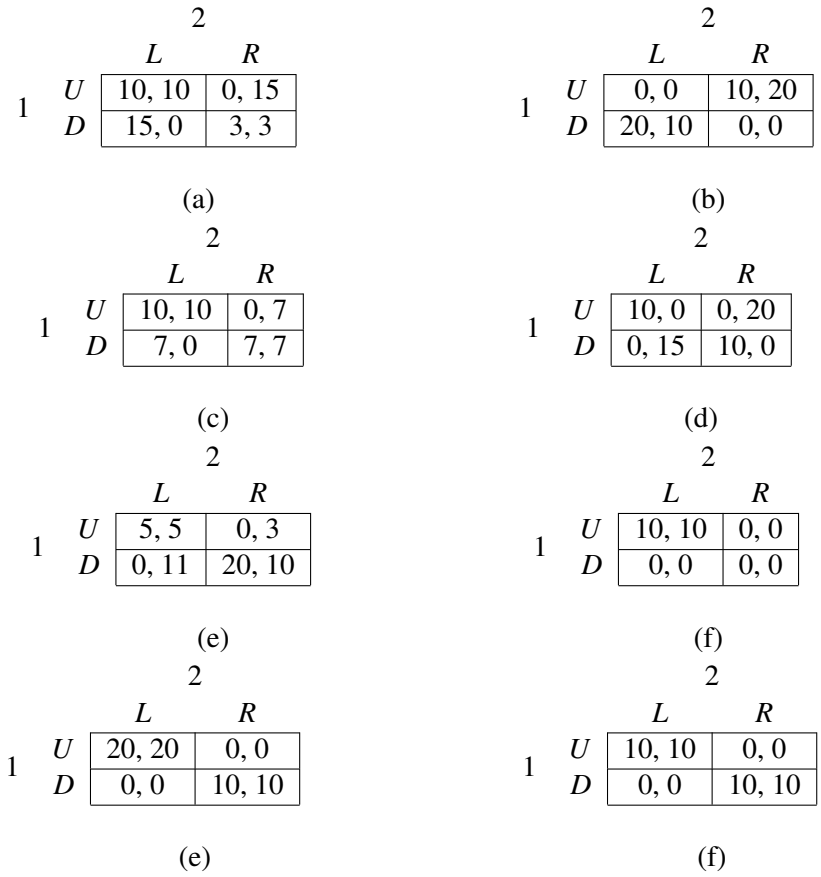
With this definition, complete preferences are neither necessary nor sufficient to make a rational choice. (In the extreme, if you are not be able to express a preference between any elements of A, any choice would be rational.) Transitivity is a sufficient condition, but not necessary. It is true, however, that if a person has preferences on a set of options A that fail transitivity for some subset of A, then this person cannot make a rational choice, in the sense defined, for this subset.[13]

Transitivity and completeness are conditions on preferences that may be more or less reasonable or interesting as positive assumptions, depending on the setting. It seems to me that "rational" in the means-end sense of choosing what leads to an outcome that you don't see as worse than some other available option is a different thing.

To this point, we are implicitly imagining very simple choice settings, where the connection between an action and the outcome in the world is basically unambiguous. For example, "would you like coffee or tea?" But much of life consists of choices where the mapping from actions to outcomes in the world is not so clear. Should I take this job offer, or keep looking? Should the leaders of a country militarily escalate an international dispute or back down? In those contexts, we often use "rational" as a normative appraisal of how a person forms beliefs about the mapping from actions to outcomes.

[13] Exercise: Give an example of complete preferences on a set of outcomes that fail transitivity for some subset of outcomes in the set but from which it is possible to make a rational choice in the sense defined here.

Figure 1: Some 2×2 normal-form games



More on this soon: How to form beliefs about what other people are going to do, who may themselves be trying to solve the same kind of problem about you, is the central normative question addressed by game theory.

9.2 2×2 games

Exercise 1. *For the 2×2 normal-form games in Figure 1, answer the following questions:*

1. *Find any and all pure-strategy Nash equilibria for each game.*
2. *For each game, are there any strategies that are never-a-best reply, or strictly dominated? (These are the same in a 2×2 game.)*

9.3 Commitment and coordination problems

We learn many things from these examples, including:

- A strategic situation can have multiple pure-strategy Nash equilibria. Multiple Nash equilibria is the defining characteristic of the class of strategic situations that we will call *coordination problems*. This is because the practical problem for the players is how to coordinate on one of the Nash equilibria.
- Nash equilibria do not always seem intuitively plausible as predictions of what might actually happen if people were to play the game. Which of the above games have Nash equilibria that seem to you like they would be bad predictions if you were to run an experiment?
- Nash equilibria are not necessarily *efficient*, or Pareto optimal. That is, it can happen that a strategic situation has multiple Nash equilibria, and one of these is better for both players than the other (examples c and f). It can also happen that there is a unique Nash equilibrium that is nonetheless inefficient: In example (a), which is known as Prisoners' Dilemma, both players would prefer (U, L) to (D, R), but (D, R) is the unique Nash equilibrium.
- In fact, in example (a), (D, R) is not only the unique Nash equilibrium. It is also the only action profile that does not involve the play of a dominated strategy. The bad outcome here (assuming, plausibly, that rational players would not choose the dominated actions U and L) has nothing to do with a failure or difficulty to coordinate, but rather should be seen as the result of *a commitment problem*. In this kind of strategic situation, the players would ideally like to be able to commit themselves to take a particular action. Here, the players would be happy to *sign a contract* saying that if 1 (2) did not choose U (L), then the player would face some punishment (like a fine) large enough to make it desirable to fulfill the agreement. If they could avail themselves of such a *third-party commitment device*, they could assure themselves the $(10, 10)$ outcome rather than $(3, 3)$.

  More broadly, we will say that a strategic situation has an aspect of a commitment problem if it has an outcome that Pareto dominates all Nash equilibria and is not a Nash equilibrium itself. Since it is not a Nash equilibrium, at least one player would have an incentive to "deviate" (because the action it calls for is not a best reply for this player), and so everyone could be made at least as well or better off if they could sign a contract to commit themselves to implementation of the non-Nash action profile.

  To be clearer:

  Definition 9. *A normal-form game* $\Gamma = \langle I, (A_i), (\succeq_i) \rangle$ **has an aspect of a commitment problem** *if there is a non-Nash action profile* $a \in A$ *(or an* α *in the set of mixed strategy combinations) that Pareto dominates every Nash equilibrium.* Γ **has an aspect of a coordination problem** *if it has more than one Nash equilibrium and there is a pure or mixed non-Nash action profile that is Pareto-dominated by at least one Nash equilibrium.*

Exercise 2. *(a) Classify the games in Figure 1 according to whether they represent commitment problems or coordination problems. Note: There is one game in the Figure that does not fit this schema, at least using the preference representation that we have at this point.*

(b) (challenging!) Construct a 3×3 *game has an aspect of a coordination problem (i.e., has multiple Nash equilibria) and an aspect of a commitment problem (i.e., has an outcome that is not Nash that Pareto dominates all of the pure-strategy Nash equilibria).*

9.4 Strictly competitive strategic situations

As discussed above, a strategic situation can have multiple Nash equilibria. However, as illustrated by game (d), a strategic situation can also have *no* pure-strategy Nash equilibria. This is typical of, but not unique to, strategic situations that are *strictly competitive* in the following sense.

Definition 10. *A two-player normal-form game* Γ *is* **strictly competitive in pure strategies** *if for all outcomes* $a, b \in A$, $a \succeq_i b$ *implies* $b \succeq_{-i} a$, *for* $i = 1, 2$.

In words (and using strict preference), if I like outcome a better than b, then you like b better than a. This also implies that there is no pair of outcomes a and b between which we have the same preference – thus *no common interests* among the pure-strategy outcomes.

We see this kind of situation often in true "games," for example, penalty kicks in soccer: The kicker wants to kick to the side the goalie is not diving towards, and the goalie wants to dive in the direction of the kick. Or, in baseball the pitcher wants to throw a type of pitch the batter does not expect, and the batter wants to correctly anticipate the type of pitch that is coming. Or, in tennis, the server wants to serve to the side of the box that is not anticipated. Or in (US) football, what play to call? Or (not a literal game), in military operations, the attacker wants to attack in a place the defender does not anticipate (or defend heavily), while the defender want to allocate forces to the sector that will get the brunt of the attack.

As we will see, if we allow for mixed strategies in addition to pure strategies, every normal-form game with finite strategy sets has at least one Nash equilibrium in pure or mixed strategies. So these strictly competitive situations do have Nash equilibria, it's just that "equilibrium" may entail that no player chooses a particular pure action with certainty. Another way to think about this is that in many strictly competitive situations, players should not be able to predict with certainty what the other will do, because the strategic logic means that one wants to be unpredictable.

So as not to seem evasive, here is a formal definition of a mixed strategy. We won't actually use these till after developing the expected-utility representation of preferences over "lotteries" on outcomes. But I occasionally refer to them before that.

Definition 11. *A* **mixed strategy** *for player* i *is a probability distribution on the set of her pure strategies,* A_i. *With* m_i *pure strategies named* $(a_i^1, a_i^2, a_i^3, \ldots, a_i^{m_i})$, *a mixed strategy is thus a list of corresponding probability numbers* $\alpha_i = (\alpha_i^1, \alpha_i^2, \alpha_i^3, \ldots, \alpha_i^{m_i})$ *giving the probability with which each pure strategy is chosen. The set of all such mixed strategies is denoted* $\Delta(A_i)$, *with typical element* $\alpha_i \in \Delta(A_i)$. *The set of all mixed strategies includes pure strategies as special cases (e.g.,* $\alpha_i = (0, 1, 0, 0)$*).*

An important class of strictly competitive strategic situations are games known as *zero-sum*, or equivalently, *constant-sum*. A normal-form game is in this class if there are two players and if for all $a \in A$,

$u_1(a) + u_2(a)$ is the same number, a constant. A famous example known as "Matching Pennies" looks like this:

		2	
		L	R
1	U	1, -1	-1, 1
	D	-1, 1	1, -1

		2	
		L	R
1	U	6, 1	1, 6
	D	1, 6	6, 1

Notice that these are strictly competitive in pure strategies and that there is no pure strategy Nash equilibrium. Notice also that with these particular payoff numbers, the players' payoffs for each outcome sum to a constant. As we will see later, after developing a way to represent preferences over actions that might lead to one thing or another, there is a unique mixed-strategy Nash equilibrium in which both players choose each of their pure actions with probability $1/2$.

Exercise 3. *Colonel Blotto. An attacker decides how to allocate forces that total $X > 0$ across $n > 1$ battlefields numbered $1, 2, \ldots, n$. At the same time a defender decides how to allocate $Y > 0$ forces to defend each battlefield. The attacker wins a battlefield if $x_i > y_i$ and loses if $x_i \leq y_i$, where these are the amounts allocated to battlefield i. A player's payoff is the number of battlefields that it wins.*

Assume that $X = Y$. Show that this constant-sum game has no pure-strategy Nash equilibrium.

Posed in the 1950s, Colonel Blotto is not just cute. Blotto-type problems show up regularly in political economy strategic situations, although to make them more tractable they are often "smoothed" by making the probability of winning a battlefield a smooth function of relative resources rather than a step function as above. We will investigate some political economy applications below. To give a sense for how they manifest, here are several examples.

- Candidates for office allocate campaign spending or spending on public services to different neighborhoods or social groups in their electoral district (Lindbeck and Weibull, 1987; Dixit and Londregan, 1995, 1996). Presidential candidates allocate campaign spending across states, competing for electoral votes. The states, neighborhoods, or social groups are like battlegrounds.
- Candidates for office allocate campaign spending to different time periods over the course of an election campaign (Acharya et al., 2020).
- A state allocates money to reduce the likelihood of successful terrorist attack across a set of possible targets, and a terrorist group then decides whether and which target to attack (Powell, 2007, 2009).
- Competing lobbying groups allocate contributions across members of a legislature, seeking to buy a majority in favor or against a bill that they care about (Groseclose and Snyder Jr, 1996).
- Protesters against the military regime in Sudan in 2018-19 start smaller protests in (quasi-) random places away from the main site to complicate and weaken the security forces' ability to repress effectively (Hassan, 2024).

- States assign missiles randomly to a large number of silos, or otherwise play "shell games" with military assets subject to remote targeting.

A technical note: $u_1(a) + u_2(a) = k$, $k \in \mathbb{R}$ and $a \in A$, has no substantive meaning when the utility functions are simply *ordinal* in the sense of just numerically representing the ordering a person's preferences over a set of outcomes. We will develop "cardinal" or "expected utility preferences" below, for which the relative spacing between the numbers corresponds to the willingness of the person to run different degrees risk in lotteries over outcomes.

If players have preferences that can be represented by expected utility functions, then we can restate the definition of a strictly competitive game in terms of pure or mixed strategies as follows: A two-player normal-form game is strictly competitive in pure or mixed strategies if for all pairs of mixed strategies $\alpha = (\alpha_1, \alpha_2)$ and $\beta = (\beta_1, \beta_2)$, $\alpha \succeq_i \beta$ implies $\beta \succeq_{-i} \alpha$ for $i = 1, 2$. A result that can be shown to follow is that a two-player game is constant-sum if and only if it is strictly competitive in pure or mixed strategies.

9.5 Commitment problems without pure-strategy Nash equilibria

But a strategic situation may not have any pure-strategy Nash equilibria even if it is *not* strictly competitive. Indeed, quite a few important and interesting social and political situations fall in this class.

For example, in a great many "principal-agent" monitoring situations, the principal wants to undertake costly monitoring if and when the agent is not working hard/not paying taxes/not complying with a law, while the agent wants to be working hard/paying the right tax/complying if she will be monitored, but perhaps not otherwise. This implies a situation in which there is no pure strategy combination that both players can expect to occur for sure, but in which it is also true that there is a dimension of common interest.

Such situations, it turns out, are commitment problems. But in contrast to our classic Prisoners' Dilemma example, in these situations the inefficient Nash equilibrium will be in mixed strategies. This is very abstract so let's illustrate with a political economy example.

Consider a Congressional district whose voters are, say, 55% Republican party identifiers and 45% Democratic party identifiers. Suppose that party activists (or the national party leaders) simultaneously decide how much money and resources to spend on mobilizing their identifiers to get out and vote in the district. For simplicity, suppose they can choose either low effort or high effort. If they choose the same effort levels or if the Republican party mobilizes while the Democrats do not, then the Republican candidate wins. If the Democrats mobilize but the Republicans do not, assume that the Democrat wins. The game below represents this situation and the natural preference ordering over the four outcomes, using $c \in (0, 1)$ for the cost of high mobilizing effort.

		Dem	
		low	*high*
Repub	*low*	$1, 0$	$0, 1-c$
	high	$1-c, 0$	$1-c, -c$

This game has no pure-strategy Nash equilibrium but is not strictly competitive (you should check to make sure you see this). As we will see later on, after we have put more structure on preferences over "lotteries," the game does have a unique mixed-strategy Nash equilibrium in which the Republican party's expected utility (or "payoff") is $1 - c$ while the Democratic party's is zero.[14] This outcome is Pareto-dominated by (low, low). The parties have a common interest here in limiting costly campaign effort, but face a commitment problem that poses an obstacle to realizing this common interest.[15]

As noted, many situations involving costly compliance with a collectively beneficial rule or law have this same basic structure. For instance, change the players so that Rep is the I.R.S. choosing whether to audit a taxpayer; a highway patrol officer choosing whether to monitor speeding; or an employer choosing whether to monitor work effort by an employee. The last is the paradigmatic case for a very large class of strategic situations that are said to involve a problem of *moral hazard*.

Definition 12. *A strategic situation involves* **moral hazard** *or is a* **moral-hazard problem** *if it is a commitment problem as defined above. That is, there is an outcome $a \in A$ that is jointly preferable to the every Nash equilibrium, but one or more players i has an incentive to deviate from a_i (i.e., a_i is not a best reply).*

The term refers to the idea that such a player i would like to promise, cross my heart, to the take the action a_i – e.g., work hard or comply with the law without being monitored – but faces an incentive structure and has preferences over outcomes that put this promise at risk.

Note that punishments are just one way to try mitigate moral hazard. Positive inducements are another. In the core literature on moral hazard in the employer-employee relationship, the employer cannot perfectly observe the employee's effort but can observe an imperfect measure of effort, such as the total sales of a salesperson. Hence, "incentive pay" (e.g., working on commission) that conditions pay on the noisy measure can partially mitigate the moral hazard problem and make both parties better off. Incentive pay will still be suboptimal, however, in so far as the employee would like to be perfectly insured against income fluctuations due to circumstances beyond her control. But this is not feasible. If the employer paid her the same amount regardless of total sales, the financial incentive to work hard is lower.

[14] In this equilibrium, the Republicans make high effort with probability $1 - c$ and the Democrats with probability c, which implies that the Republican candidate wins with probability $1 - c^2$.

[15] Since the Democratic party is not strictly better off at (low, low) versus the mixed-strategy Nash equilibrium, and we think of the parties as having opposed electoral interests so that the Democrats might well *like it* that the Republicans need to spend effort to keep their safe seat, we might want to modify the payoffs to reflect this, which could remove the element of common interest. But this is imagining the larger game of how to allocate campaign money across multiple districts. In that larger game, you can easily see how, if there are similarly safe districts for the Democrats, the two parties could have an interest in an agreement to not spend of lot mobilizing effort in each others' safe districts. When we get to the theory of repeated games, we will study how such agreements might be rendered stable in the sense of Nash equilibrium, despite the "short-run" incentives to choose *high* seen in the 2×2 game here.

Essentially the same issue arises in the principal-agent relationship between voters and a politician who is supposed to represent them in the legislature or other office. The politician might want to commit to not be corrupt or not to cater to special interests that make large campaign donations if, by doing so, this would make typical voters more supportive at the polls. But so much of what politicians do is not directly observable by voters (aka "hidden actions") that there is a major, indeed fundamental, moral hazard problem in electoral accountability relationships. Like the employer above, voters have to use noisy signals of politicians' actions, such as how the economy is doing, reports of scandals, and so on. More on electoral accountability models later in the book.

9.6 Types of coordination problems

Figure 1 depicts several different examples of coordination problems, which have the defining feature of multiple Nash equilibria. But they are not all the same. They help to illustrate three kinds of coordination problem that appear constantly in more developed models in more specific contexts.

1. *Pure coordination*. The players don't care *how* they coordinate, they just want to coordinate to avoid outcomes that they both agree are bad. For example: We want to meet for coffee, but which Coupa? Our phone call dropped, who should call back? Move left or right when trying to pass on the sidewalk? Should the sounds that we use to refer to the thing in the world that is a barking pet be "dog" or "chien" or "mbwa" or "urzak" or ...? Should all assume that Green means go and Red means stop, or that Red means go and Green means stop? Should we all drive on the left side of the road, or the right side? What should electrical plugs, or railroad gages, or internet protocols [or all kinds of standards] be, exactly?

   In the 2×2 representation (say):

		2	
		L	R
1	U	10, 10	0, 0
	D	0, 0	10, 10

2. *Bargaining problems*, or in the 2×2 form often called *Battle of the Sexes*. The players want to coordinate to avoid outcomes they both dislike relative to coordinating, but they have conflicting preferences over how to coordinate. Thomas Schelling and, arguably, John Nash, saw almost all of human interaction – in politics, economics, and everyday interactions – as being bargaining problems in this sense.

   A 2×2 representation of Battle of the Sexes is:

		2	
		L	R
1	U	0, 0	1, 2
	D	2, 1	0, 0

Exercise 4. *Two players, $i = 1, 2$, simultaneously write down a number $a_i \in [0, 1]$. If $a_1 + a_2 \leq 1$, then player i gets share a_i of a monetary prize, e.g., \$100. If $a_1 + a_2 > 1$, then neither player gets anything. Assume they like more money rather than less.*

(a) Find any and all pure-strategy Nash equilibria.

(b) Are there any inefficient equilibria?

(c) Compare this **Nash demand game** *to Battle of the Sexes and to Pure Coordination.*

(d) If you were to play this game with another student in the class, what would you write down, and why?

Bargaining problems are central to all politics, and to human life more broadly. In what follows, they get their own sections both for normal- and extensive-form treatments.

3. *Assurance games* or *Stag Hunt*. In these, there is a Nash equilibrium that Pareto dominates all other Nash equilibria. You might think that it should then be obvious to rational players that they should focus their attention on this jointly best Nash equilibrium, and expect others to choose the relevant action that will bring this about. And in some variants of the Stag Hunt problem, this is entirely correct and experimentally supported. For instance,

		2	
		L	*R*
1	*U*	20, 20	0, 0
	D	0, 0	10, 10

But consider the following strategic situation.

Exercise 5. *There are $n > 1$ players, and you are one of them. Every player chooses whether to write down 10 or 20, thus $A_i = \{10, 20\}$. If a player writes down 10, this player gets \$10 no matter what anyone else writes down. If all players write down 20, then every player gets \$20. If a player writes down 20 and any other player writes down 10, the player who chose 20 gets \$0.*

(a) What would you write down if $n = 25$? 4? 100? If your answers differ, why is that?

(b) Find any and all pure-strategy Nash equilibria of this game (for arbitrary n).

This n-player game is strategically similar to the following two-player 2×2 game, which also has an action for each player that would clearly be both collectively and individually (in terms of best replies) optimal if everyone chose it. But the fear that other players might not "act rationally" by choosing the good action is, in a Nash-equilibrium sense, self-confirming: Choosing the good action is risky and dangerous if the other players don't also choose it, while choosing the other action is safer. And we don't even need to assume players who actually are "irrational" in that they can't see what is best, or who perversely don't want to act on it for some reason. What if

the other players *are* able to see what the best thing to do would ideally be, but they worry that *you* might not? What if you worry that they worry that you worry? Etc.[16]

		2	
		L	R
1	U	30, 30	0, 20
	D	20, 0	20, 20

Where in the real world do we see strategic situations that are structured like Stag Hunt games? Quite often in settings where there are many agents ($n \gg 2$) who choose, essentially, between two actions, call them $A_i = \{0, 1\}$. Some examples:

- Players are people deciding whether to go out to protest against the government ($a_i = 1$) or stay home ($a_i = 0$). The more who protest the greater the chance that the regime falls, but if you protest by yourself or with very few others, you just get locked up or worse. Protesting without getting locked up feels good, so it is not a large Prisoners' Dilemma problem if enough people hate the regime.
- Crashes in asset markets: $a_i = 0$ means try to withdraw your money from the bank, or sell your stock holdings, or sell the currency. $a_i = 1$ means don't do this. This basic structure is common in models of bank runs, currency collapse, stock market crashes, market "bubbles," and so on. Some Keynesian-style theories about recessions and depressions have this flavor also: "Consumer confidence" is like a Stag Hunt situation, or business investment depends positively on expectations about how much other businesses will be investing and how much consumers will be buying, which depends in part on aggregate business investment.
- Some dysfunctional cultural patterns: Maybe we would all prefer it if everyone was generally nicer and more courteous in daily face-to-face or online interactions, and most people would also want to reciprocate niceness with niceness. But if the typical best reply, emotionally, to a low standard of niceness is also to be not-so-nice, then we may collectively get stuck in an unpleasant equilibrium illustrated by the "10" equilibrium in the game above.

  In online fora it often seems that at first most people are nice, not jerks, but a small number of jerks causes some normal people to opt out. This increases the proportion of jerks which drives out more normal people, leading to a nearly pure troll equilibrium.

  Or consider a social norm like queuing. If everyone expects everyone else to queue and to sanction queue-jumpers, they may be all better off than if everyone expects everyone else to try push and fight their way to be the first one served, or to get on the bus, etc. Note how in each equilibrium queuing or struggling to be served is a best response to what others are expected to do.

[16]The label Stag Hunt comes from an example Rousseau deploys in *Discourse on Inequality*, in which he has a group of hunters choosing whether to hang together, which is necessary if they are going to catch a stag, or individually run off to grab a passing rabbit. My impression is that Rousseau was actually trying to say something about impatience, or time preferences, or shortsightedness, but I'm not sure (it is also seems possible that he was trying to describe a Prisoners' Dilemma situation, but didn't quite have it). Regardless, it's a good story for illustrating the important strategic dynamics of what we now call Stag Hunt-type problems.

Or, consider low-level government officials choosing whether to ask for a bribe to provide a service to a citizen ($a_i = 1$) or not ask for a bribe ($a_i = 0$). If no one else is demanding bribes, then it may be risky to demand one, as you stand out for violating the norm. But if everyone is demanding bribes, you want to as well because you can get away with it. Notice that here, what is taken as collectively good for bureaucrats (all demanding bribes) is collectively bad for citizens. (Alternatively, the "everyone asks for a bribe" equilibriium could be bad for the bureaucrats as well if we consider aggregate effects on the economy, or the fact that they also may have to pay lots of bribes.)

- Political order: In *Leviathan*, Thomas Hobbes' story about escape from the "condition of Meer Nature" (anarchy, the State of Nature) depicts the problem in Stag Hunt terms. Hobbes says that if others are willing to subject themselves to the authority of a sovereign, then rationally you should do this also, because life under even a crappy sovereign is so much better than life in anarchy ("nasty, brutish, and short ..."). But he also says that if others are not choosing to subject themselves to Leviathan (the state, or the ruler), then it is rational and justified to do all the things in the State of Nature to preserve your own life, even though by doing so everyone is contributing to this being a nasty state of affairs.
- Likewise: Let $a_i = 1$ mean that i pays taxes and/or complies with the law, and $a_i = 0$ means that i does not pay taxes or comply with the law. If no one else is paying taxes, the government does not have the resources to enforce tax collection or to provide basic public goods, in which case it may not be a best reply for you to pay taxes or comply with the law. But if most or all other people *are* paying taxes and complying with the law, then the state is strong enough to conduct audits and to prosecute non-payers and non-compliers, in which case it can be a best reply to pay and comply yourself. So there can be multiple Pareto-ranked Nash equilibria.

  Notice that in this account the phenomena of *political power*, *political authority*, and *political order* – all central concepts in Political Science – are understood as self-enforcing and self-confirming social equilibria. For individuals, the phenomenon appears as an external fact about the social world, but at the same time each individual's choices are producing and reproducing the phenomenon. As Hobbes put it, expressing an equilibrium logic, "Reputation of power, is Power; because it draweth with it the adhaerence of those that need of protection" (followers, supporters).

9.7 A more developed example: Revolution

The following example develops a richer model to explore implications of the idea that revolutions (and coups, and bank runs, etc.) are like Stag Hunt coordination problems, strategically.

Society, or residents of a capitol city, are represented as a continuum with unit mass, with individuals denoted by $i \in [0, 1]$. Person i has value $v_i \in \mathbb{R}$ for overthrow of the current regime in power, and value (normalized to) zero for the status quo. Thus, a person with $v_i < 0$ prefers the current regime to overthrow, and a person with $v_i > 0$ prefers overthrow. The distribution of v_i in the population is given by the cumulative distribution function $F(v)$, which means that $F(v)$ is the share of society that has $v_i < v$.

Individuals simultaneously decide whether to go out to protest against the regime in the main square. Represent this as action $a_i \in \{0, 1\}$, where zero is staying home, and 1 is going out to protest. Let $s \in [0, 1]$ be the proportion of the population that protests, which is an endogenous quantity that we will solve for in equilibrium. For simplicity, suppose that the probability that the regime falls is equal to s.[17] We normalize the value for staying home – not protesting – to be zero.

If i protests, i's payoff is $v_i + \epsilon v_i$ if the regime falls and $\epsilon v_i - k$ if the regime survives.

Here, $\epsilon \geq 0$ is a small number that represents *the expressive utility of protesting*. If an individual dislikes the regime, so that $v_i > 0$, notice that such a person gets a psychological benefit ϵv_i from protesting whether or not the protests succeed in ending the regime. This psychological benefit is greater the more the person dislikes the regime (higher v_i). By contrast, a regime supporter ($v_i < 0$) dislikes protesting per se, independent of the outcome (relative to staying home).

$k > 0$ is the *expected cost of being arrested and punished* for protesting if the regime survives. We could make this depend on the size of the protests – for example, the expected cost could be $k(1 - s)$. But for starters it is clearer and simpler if we don't.

Putting things together, we have defined a normal-form game with the following expected payoffs:

$$u_i(a_i = 0, s) = sv_i + (1 - s)0 = sv_i$$
$$u_i(a_i = 1, s) = s(1 + \epsilon)v_i + (1 - s)(\epsilon v_i - k) = \epsilon v_i + sv_i - (1 - s)k.$$

Notice that an individual gets the same expected payoff concerning regime change regardless of whether she protests (sv_i). This is because we are assuming individuals are "very small" and their individual presence at the protest has a negligible impact on the probability of success, by itself. You can easily see, then, that if there was no "expressive utility" for protesting, then staying home would be a weakly dominant strategy – for any $s < 1$, staying home avoids the costs but gets the same benefits, for a person with v_i. In other words, in the case of $\epsilon = 0$, there is a strong incentive to *free ride*, which makes for no collective action and no protesting, even when everyone hates the regime.

But if people who dislike the regime take some direct pleasure in expressing their principles, then some people may prefer to protest, depending on their beliefs about what others are doing. Given the expectation that s share will protest, i prefers to protest if

$$\epsilon v_i + sv_i - (1 - s)k \geq sv_i$$
$$v_i \geq (1 - s)k/\epsilon.$$

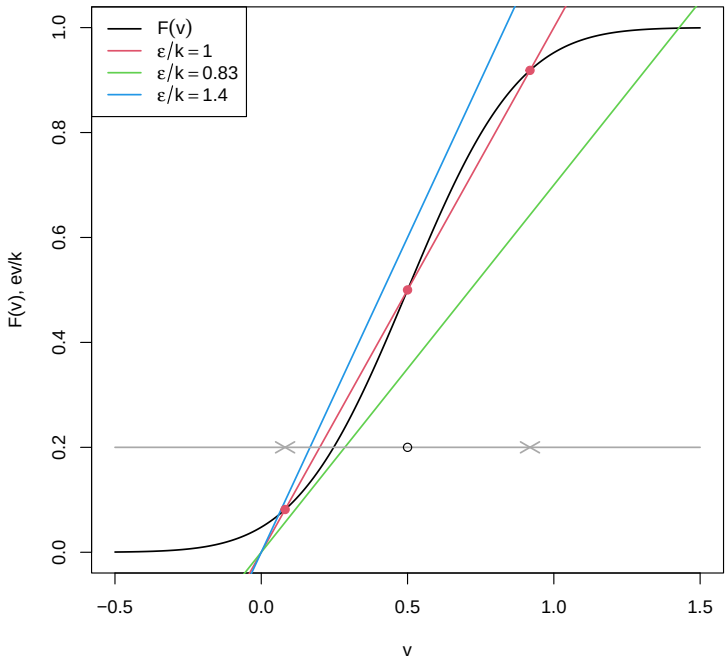
Those who dislike the regime the most protest, and the threshold type is $\hat{v} = (1 - s)k/\epsilon$. This type is decreasing (less extreme dislike of the regime) the more others are expected to protest (larger s). Thus there is a *strategic complementarity* in protest behavior in this model with expressive utility: The more a $v_i > 0$ type thinks others are going to protest, the greater her payoff from protesting.[18]

Observe also that for types who like the regime, $v_i < 0$, staying home is a dominant strategy, a best

[17] We could have this be $p(s)$ where $p' > 0$.

[18] This is in contrast to situations with *strategic substitutes*: The more others are doing X, the less you want to do X. This can be true in the free riding case, for example, or in Prisoners' Dilemma-type situations more generally.

Figure 2: Equilibria in the rebellion game



reply no matter what anyone else does. Thus there will always be at least $F(0)$ share of the population staying home.

We have not yet solved this game for Nash equilibria. We have only defined i's best reply function (what to do for each s, which summarizes what everyone else is doing). A Nash equilibrium is a set of strategies for each i such that each i is choosing a best reply to what everyone is in fact choosing. How to find this?

We know that in any equilibrium, all types v_i greater than a cutpoint value, $\hat{v} = (1 - s)k/\epsilon$, will be protesting. Thus in equilibrium – when expectations are correct so that everyone correctly anticipates the size of the protest and chooses accordingly for themselves – the s in this expression will equal $1 - F(\hat{v})$, the share of people with $v_i > \hat{v}$. So we have the equilibrium condition

$$\hat{v} = (1 - s)k/\epsilon = F(\hat{v})k/\epsilon$$
$$\epsilon\hat{v}/k = F(\hat{v}).$$

So equilibria are wherever the linear function of v on left-hand side intersects with the cdf of v. Figure 2 illustrates using a Normal distribution for values for regime overthrow, centered at $v = .5$ with standard deviation .3, for three different ratios of expressive utility to expected cost of arrest, ϵ/k. Keep in mind that $F(v)$ on the y-axis is the share of people *staying home*, and the smaller is $\hat{v}$, the bigger is the size of the equilibrium protest turnout, $1 - F(\hat{v})$.

Notice that there can be either one or (generically) three Nash equilibria (indicated by the solid points for the case with three), depending on the exogenous factors, which are ϵ, k, and the shape and location of the distribution of values for overthrowing the regime, F.

When average regime discontent (mean of F) is low, the S-curve is shifted towards the left relative to $v\epsilon/k$, and the only intersection will be at single high value for $F(\hat{v})$. This means that almost everyone stays home (where green and black lines met). If the mean discontent is quite high, on the other hand, the only intersection will where almost everyone turns out (where the blue and black lines meet). These are cases of unique equilibria with low and high mobilization, respectively.

In between, there will be a range of average societal discontent such that there are multiple equilibria, as shown, and generically there will be three if there is more than one – low and high turnout equilibria, and an intermediate one. Using a related set up, Kuran (1991) argued on this basis that a very small shift in underlying preferences in the population could suddenly and surprisingly make a situation prone to revolution (all of a sudden there are high-mobilization equilibria).

We can also imagine keeping the distribution of values for the regime fixed, and consider the impact of change in expressive value for protesting or costs of protesting. Expressive value for protest might be "shocked" upwards by public anger in response to a case of regime brutality or other malfeasance that attracts coordinated focus. Or, expected costs for protest might be shocked downwards by news that, say, a major-power supporter of the regime was withdrawing its support, or a domestic political crisis such as elite fighting that makes the police divided or reluctant to obey orders. In the model, such changes increase the slope of $v\epsilon/k$, and accordingly increase equilibrium turnout from small numbers to, potentially, a sudden "tip" to mass protest.

Consider a situation with multiple equilibria. You might well be wondering how would it be possible that a massive group of people would, in the novel circumstances of a possible rebellion, coordinate on common expectations about the likely size of the protest? Good question. To some degree, the next section below will speak to this issue of *equilibrium selection* by introducing the idea of focal points and evolutionary stability. For now, consider the following more heuristic argument that relates closely to Thomas Schelling's work on *tipping games* (Schelling, 1978).

Let's imagine that the people in our model face this decision problem of whether to go protest or not on successive days. Suppose that they follow a simple, myopic (not fully rational) decision rule: Go protest today if s_{t-1}, the size of the protest yesterday, was large enough that given one's v_i, protesting is a best reply (i.e., best reply to yesterday's s). For the first day, $t = 0$, let's just stipulate that a random share of the population, say people with $v_i \geq \hat{v}_0$ for some $\hat{v}_0$, go out to protest. (Perhaps these decisions result from people observing a common signal like a brutal event, and making guesses about collective anger this will occasion.)

Then on day $t = 1$, $\hat{v}_1 = F(\hat{v}_0)k/\epsilon$, and we have a dynamic process $\hat{v}_t = F(\hat{v}_{t-1})k/\epsilon$. The size of the protest on day t is $s_t = 1 - F(\hat{v}_t)$, so the protest size is increasing when

$$\begin{aligned} \hat{v}_t &< \hat{v}_{t-1} \\ F(\hat{v}_{t-1})k/\epsilon &< \hat{v}_{t-1} \\ F(\hat{v}_{t-1}) &< \hat{v}_{t-1}\epsilon/k, \end{aligned}$$

and thus whenever the cdf in Figure 2 is below the line $\epsilon v/k$. The protest size is decreasing when the reverse is true. The gray arrows in the Figure indicate the direction of $\hat{v}_t$ from any initial starting point $\hat{v}_0$.

Consider the case of three equilibria, labelling them $\hat{v}^h$, $\hat{v}^m$, and $\hat{v}^l$, where $v^h < v^m < v^l$. The superscripts stand for *turnout* levels – high, medium, and low. (Recall that lower $\hat{v}$ means that more people turnout, because more people have a v greater than $\hat{v}$.)

You can see that the high and low turnout equilibrium levels are "attractors" for initial turnout levels $v_0 < v^m$ and $v_0 > v^m$, respectively. From any initial turnout v_0 different from an equilibrium point, successive days will see the crowds either grow or shrink towards $1 - F(v^h)$ or $1 - F(v^l)$. And v^m is "dynamically unstable" in the sense that any small perturbation from it leads to escalation or decrease towards one of the other equilibria.

There is a parallel and close connection to the simple 2×2 Stag Hunt games such as Figure 1(c), which have two pure-strategy Nash equilibria, one of which Pareto dominates the other. As we will see, these also have an intermediate mixed-strategy equilibrium that is arguably "unstable" in a sense similar to v^m here.

10 Correlated equilibria and focal principles

Here is a simple representation of the strategic problem faced by two car drivers who arrive at a four-way intersection with stop signs.

		Driver 2	
		Wait	Go
Driver 1	Wait	play again after slight delay	good outcome, slight delay for 1
	Go	good outcome, slight delay for 2	possible crash

This strategic situation is similar to the coordination problem Battle of the Sexes.[19] There are two pure-strategy Nash equilibria, $\{Go, Wait\}$ and $\{Wait, Go\}$, both preferable to failure to coordinate, but with drivers having a slight preference for getting to Go first.

But how to coordinate? In real life, drivers have a large set of options for solving the problem of how to select one of the two pure strategy equilibria. All of them involve using some feature of the specific context to *break the symmetry*, to "designate" one of the two as the driver who should go first. Means

[19] It is not the same because if both wait, they "play again" after the delay, and keep playing until someone goes. This is really suggesting an extensive-form model with periods $t = 1, 2, 3, \ldots$, where the game ends and payoffs are received in the first period that a player chooses *Go*. We will formally analyze this game later in the course.

of breaking the symmetry include, first of all, the rules of the road: Which driver arrived first, or if at approximately the same time, who is to the right of the other? Second, in practice, the symmetry may be broken by one of the drivers being more aggressive about starting to go, or clarity about which road is busier. Third, some intersections are "governed" by traffic lights rather than stop signs.

Notice that no such specific contextual features are explicitly encoded in our abstract formulation of a normal-form game Γ. But as Thomas Schelling stressed, it is completely obvious that people constantly use specific contextual features and mutual knowledge (or guesses about what they mean for behavior) to solve coordination problems.[20]

In a static version of the intersection problem, if the drivers do not have any feature available that would allow them to break the symmetry, then the probability of failure to coordinate has to be at least 1/2. Consider the following Battle of the Sexes, and suppose that each chooses Go with probability p. (They cannot choose with different probabilities if there is no way to assign which probability to which driver.) Then they fail to coordinate with probability $(1-p)^2 + p^2 = 1 - 2p + 2p^2$, which is minimized at $p = 1/2$, so that coordination and failure to coordinate are equally likely. That is really bad, much worse than if they could condition on some common signal that would allow them to *correlate* their action choices.

		2	
		Wait	*Go*
1	*Wait*	0, 0	20, 21
	Go	21, 20	0, 0

	$1-p$	p
$1-p$	$(1-p)^2$	$(1-p)p$
p	$p(1-p)$	p^2

With a traffic light (and neglecting Yellow and low probability mistakes), they can implement the following ex ante probability distribution on outcomes. Here, both drivers choose the action "Go on Green, Stop (Wait) on Red," and given that each expects the other to do this, go on green and stop on red is a best reply. Observe how this probability distribution on outcomes is not possible with independent action choices.

	Wait	*Go*
Wait	0	1/2
Go	1/2	0

This is an example of a *correlated equilibrium*, a solution concept that considers a normal-form game Γ extended by adding signals on which players can condition their action choices.

Some game theorists argue that correlated equilibrium is preferable to Nash equilibrium as a solution concept, on the grounds that real-world strategic situations virtually always come with lots of specific context that players can in principle use to correlate their action choices.[21] For example, Nash equilibrium rules out the quite plausible seeming "traffic light" 50-50 distribution in Battle of the Sexes as a

[20] See especially Schelling, *The Strategy of Conflict* (Harvard University Press, 1960), chapter 3.

[21] See especially Gintis (2014).

prediction, whereas correlated equilibrium allows it. Note that if the signals (i.e., context) are public, so that each knows what the other is conditioning on, then the action choices in a correlated equilibrium will themselves form a Nash equilibrium in Γ. It's just that now the context is giving them a way to *select* a Nash equilibrium that they could not implement with independent strategies.

To formalize this idea, we enrich our description of a strategic situation by adding *signals* that the players can condition their action choices on. Let M_i be a finite set of signals (or "messages") that i might observe prior to choosing action $a_i \in A_i$. A strategic situation will now be characterized by Γ plus the set of all message combinations, $M = \times_i M_i$, along with a commonly known prior probability distribution on the messages, $\mu\colon M \to [0, 1]$. Thus, in a two-player game, $\mu(m_1, m_2)$ would be the probability that the players observe signals $m = (m_1, m_2)$. Given the prior distribution and having observed her own signal, player i can form an updated belief about the what player $j \neq i$ saw. In the case of *public signals*, both observe $m = (m_1, m_2)$, so this posterior distribution just puts probability 1 on what the other player actually saw. Public signals is the natural assumption for situations like the traffic light example above.[22]

Definition 13. *With publicly observed signals $m \in M$, strategies $a_i^*\colon M \to A_i$ form a* **correlated equilibrium** *if, for all i and for all m that may occur ($\mu(m) > 0$), $a^*(m) \succeq_i (a_i, a_{-i}^*(m))$ for all $a_i \in A_i$.*

For publicly observable signals, all we are really doing here is pointing out that in real-world strategic situations, players can normally condition their actions on observable features of the specific context. This may have no value in strategic situations like Prisoners' Dilemma, where they have dominant strategies, but is highly relevant for coordination problems, where precisely the point or challenge is to guess what the other thinks you are going to do, and vice versa and so on.

Exercise 6. *Imagine that you are player 1 in the two games depicted below. You are going to be paired at random with some other player, who will have made a choice as player 2. Imagine that the payoffs represent dollars. Make a choice for each game, and explain your choice.*

		2: *A*	2: *B*
1	*A*	10, 10	0, 0
	B	0, 0	10, 10

		2: *B*	2: *A*
1	*B*	10, 10	0, 0
	A	0, 0	10, 10

Normally, most people choose A in the first game, so that a large fraction of players successfully coordinate on (A, A). In the second game, typically more people choose B but many also choose A, so that there is a much higher rate of coordination failure when they are paired together.

This is an example of how people use *focal points* or *focal principles* to solve the problem of equilibrium selection in a coordination game. The first game is relatively "easy" because

1. "A" is the first letter of the alphabet, and things that are commonly known to be first according to some natural metric have a particular salience – there is only one first letter, one first positive number, etc., but there are many non-first letters.

[22] Private signals ...

2. The (A, A) outcome is in the top-left box, and in English we read pages starting at the top-left. So this is a second way that (A, A) comes first, in a sense, and so is salient by this second principle as well.

In the second game, these two focal principles are in conflict – (B, B) now "wins" by the top-left focal principle – which accounts for the greater difficulty for players to coordinate.

Thomas Schelling noticed that when confronted with coordination problems, people study the specific context and try to guess what focal point or principles the other players will apply to select, or focus on, one of the equilibria. Using examples from ingenious games he played with his students, players did much, much better by anticipating focal points than a computer or any "dumb machine" that just has access to the formal description of the game (absent context, or M from above) could possibly do.

Schelling pointed out that this means that the whole program of asking "What is it rational to do in a strategic situation?" by stripping away context to get to an abstract formulation like Γ is self-contradictory, or misguided in a way. For coordination problems, what we learn is that it would be irrational to strip away the context to try to figure out what to do! Since the whole program of microeconomic theory consists of proposing models that strip away all but the (proposed) strategic crux of the matter, this is a fundamental critique. At a minimum, we could say that Schelling's critique suggests that correlated equilibrium is a more empirically plausible solution concept than Nash equilibrium.[23]

When I was a kid we visited both sets of grandparents on holidays and some other occasions. With my father's parents, there would normally be snacks and drinks before dinner, and it was understood that the women would prepare and serve the snacks, possibly assisted by children. Drinks with alcohol were prepared and served by my grandfather, usually.

Everyone wants some snacks and drinks but it doesn't make sense for everyone to prepare them. So this problem has basic features of a Battle of the Sexes strategic situation. The coordination problem was solved by conditioning actions on gender and age, in a manner that reflected and enacted a cultural norm. In so far as "serving" is understood to express deference and hierarchy, the cocktail-hour coordination problem was solved according to a general focal principle that puts men at the top of a social hierarchy.[24]

A great deal of politics – "identity politics" in particular – involves contestation over social norms and roles governing day-to-day coordination problems.

[23] Schelling (1960, 98): "The assertion here is not that people simply *are* affected by symbolic details but that they *should* be for the purpose of correct play. A normative theory must produce strategies that are at least as good as what people can do without them. More, it must not deny or expunge details of the game that can demonstrably benefit two or more players and that the players, consequently, should not expunge or ignore in their mutual interest."

[24] Alcohol is dangerous and manly, thus preparing and serving it also reflects the hierarchy? Interestingly, my mother's parents did not drink alcohol and at their house, women and children also prepared the drinks – water, tomato juice (yuck), and occasionally soda, which my grandfather quietly disapproved of ("sugar water").

11 Stronger notions of stability

Here is game (f) from Figure 1, now given symmetric action labels.

		2	
		A	B
1	A	10, 10	0, 0
	B	0, 0	0, 0

You saw that this is a coordination game with two pure strategy Nash equilibria, (A, A) and (B, B). (B, B) is Nash because if you expect the other player to choose B, you are going to get zero whether you choose A or B, so it just doesn't matter and B is a best reply. This is true for both players so we have a Nash equilibrium.

But you probably also observed that (B, B) does not seem at all plausible as a prediction of how people would choose in this strategic situation. A point: Although rational players cannot mutually expect a non-Nash action profile, some Nash equilibria can seem like better predictions than others.

Why do we expect that a rational person would not choose B? Here are two arguments.

1. (A, A) is not only Nash but is also clearly better for both players than any other outcome, which makes it *focal* in the sense of the last section.
2. If you think that there is even a tiny positive probability that the other player will choose A, then B is no longer a best reply. No matter what the other player's strategy is, you are always at least as well off or better off if you choose A than B.

11.1 Weak dominance

The second argument expresses a new generalizable candidate for "a bad strategy": An action that, no matter what other players choose, is at least as bad or strictly worse than some other available action. (Compare to the definition of strict dominance.)

Definition 14. a_i' **weakly dominates** a_i *if* $(a_i', a_{-i}) \succeq_i (a_i, a_{-i})$ *for all* $a_{-i} \in A_{-i}$, *and* $(a_i', a_{-i}) \succ_i (a_i, a_{-i})$ *for some* a_{-i}. *We also say, in this case, that* a_i *is* **weakly dominated**.

One proposal for *refining* the set of Nash equilibrium – also known as an *equilibrium refinement* – that is frequently used in applied work is to restrict attention to Nash equilibria that do not involve the play of any weakly dominated strategies. Obviously, this refinement eliminates or rules out the (B, B) Nash equilibrium in the game above.

Exercise 7. *Consider a majority-rule election between a Republican and a Democratic party candidate, where the players are n voters, n very large (and odd, for convenience). Suppose that* $r > (n+1)/2$

voters are Republican-party identifiers and the rest are Democrats. Voters simultaneously vote for one or the other, $A_i = \{R, D\}$. Voters care only about the party of the winner, preferring their co-partisan. Voters are assumed to have no direct or "expressive" preference over how they vote, they care only about the outcome (so voting is assumed to be purely instrumental).

(a) Characterize the full set of pure-strategy Nash equilibria in this voting game. (You can do this in words rather than with a lot notation, if you want.)

(b) Which Nash equilibria involve the play of weakly dominated strategies? Give an example and explain why one of the voter's votes is weakly dominated (that is, how it satisfies the definition).

(c) Which Nash equilibria/um does not involve the play of any weakly dominated strategies?

In "one-shot" (normal-form) strategic situations with a large number of players, ruling out weakly dominated strategies is usually a compelling refinement of Nash. This is because in a large population, people will anticipate at least a small amount of variation in other peoples' choices. The refinement is often compelling in two- or small-number-of-players games also, as in the example above, but not necessarily. Consider the following example. What would you do if you were player 2? Player 1? In this case, the focal gravity of (U, L) works against the force of weak dominance for player 2.

		2	
		L	R
1	U	10, 10	0, 10
	D	0, 0	1, 1

11.2 Expressive utility and a first model of democratic failure

This next exercise is not strictly about weakly dominated strategies, but in my view it does illustrate how the previous exercise – and standard arguments about electoral democracy in general – overstate the case for efficiency of majority rule even with just two alternatives.

Exercise 8. *The region of Freedonia is holding a referendum on possible secession and independence from the United Realm. Freedonia has n voters (again, remarkably, n is odd). All voters realize that if a majority votes for secession, independence will happen, and there will be significant economic costs for every Freedonian. To be concrete, suppose that every voter's annual income will decline by $1,000. However, Freedonians are all nationalists, so that they find it slightly distasteful to vote against independence. Suppose we represent this distaste as worth $\$\epsilon > 0$, where ϵ can be very small.*

Let $A_i = \{0, 1\}$, where $a_i = 0$ is a vote against and $a_i = 1$ is a vote for independence. Let $x = 1$ if a majority votes for independence, and $x = 0$ if the referendum fails. Thus, voter i's payoff (preferences) can be represented here as $-1000x + \epsilon a_i$.

1. *Characterize all pure-strategy Nash equilibria. "Verbally" is fine, if you give the argument for why each is indeed Nash.*

2. *Which are efficient? Which inefficient? Which do you see as most plausible and why?*

3. *What kind of strategic situation is this?*

In this example, voting is said to have *expressive utility* or value – people care not only about the outcome, but also about the act, or how they vote independent of the outcome. The example illustrates that for democracy to work as it is intended to, a population needs to have a good *culture of voting* (or, norms of voting). People should believe that they *should* vote as if they were decisive. In other words, any small satisfaction one gets from the act of voting should come from voting as if one is making the decision for the polity. Otherwise, individual failures to take account into the "externalities" of expressive voting can at least in principle make for collectively very bad outcomes.

You will have noticed that the efficient equilibria in this voting game require a truly heroic and wildly implausible degree of coordination among voters. In these equilibria, it is commonly and correctly expected that the referendum will fail by exactly one vote, and each voter knows whether to be a Yes or No voter.

In my opinion, in this example and in many other models of voter behavior in mass elections, it makes far more sense to model voters as a continuum, so that the electorate is infinitely large. With an infinite number of voters, no voter's vote can be decisive, ever. For the case at hand, if we change the model so that the electorate is represented by a continuum, then the *unique* Nash equilibrium has every voter voting for secession. $a_i = 1$ becomes a dominant strategy, and the strategic problem is a large-scale Prisoners' Dilemma (a collection action problem). Voters fail to internalize the negative externality (cost) that their Yes vote imposes on others. As result everyone is worse off than they would be if the referendum simply wasn't held.

In the real world, there is essentially zero chance that an individual voter can be decisive in a large election. As a result, *instrumental rationality simply does not pin down how a person should rationally vote*. We may want to assume in our analyses that voters vote as if they imagine themselves to be decisive, and thus choosing the outcome for everyone, self included. But we should recognize that if they do so this is a matter of cultural coordination on a social norm, not a clear-cut implication of instrumental rationality in the same sense that, say, I should not order steak if I am a vegetarian.[25]

11.3 Evolutionarily stable strategies

Returning to the idea of ruling out weakly dominated strategies: Another way to express the argument about (B, B) in the game above is to say that it shows that as a stability criterion, Nash equilibrium is *too weak*. It permits action profiles that we have reason to think would not in fact be "stable."

But stable vis-a-vis what? One important class of answers comes from evolutionary game theory that

[25] Acharya and Meirowitz (2017) make a parallel point concerning models of large elections in which voters are assumed to vote based on the infinitesimal chance that they are decisive, but they are decisive only if an unknown "state of the world" is such that they would want to vote strategically against their own private information about the state. They point out that introducing even a vanishingly small amount of (what is in effect) noise makes it rational for voters to vote sincerely again, which in this kind of model restores efficiency of majority rule for aggregating voters' private information about what the best policy is.

was developed by biologists. Their answer, formalized in the concept of *evolutionarily stable strategies* or ESS, is "stable against invasion by mutations."

To illustrate, suppose that a large population of a certain kind of insect can have two phenotypes, X and Y. Both X and Y emit a pheromone that tells others of the same species that they are present and can mate, but phenotype X has developed a pheromone variant that is detectable over a larger range by other X types (and not by Y types). The matrix shows the "fitness" of each phenotype when paired, spatially, with another insect. Fitness is a number positively related to reproductive success, for example, it is proportional to expected number of surviving offspring.

		2	
		X	Y
1	X	3, 3	2, 2
	Y	2, 2	2, 2

If a population of these insects starts with almost all of them as Y types, over time it will be successfully "invaded" by X types because their average fitness is slightly higher. X's do just as well when (randomly) matched with Y's, and slightly better on the initially rare occasions when they are matched (in the same general location as) another X-type.

To formalize and develop this idea, consider a symmetric, two-player game, meaning that both player's have the same strategies available, $A_1 = A_2$. We will imagine a large population that is randomly paired off in successive periods. (In social science contexts this is referred to as a *social matching game*.) In fact, at the risk of some confusion, for this model let's call the set of actions/phenotypes available to a player A, so that the set of *outcomes* is now $A \times A$.

Let $u(a, b)$ be a player's payoff, or fitness, if a player "choosing" $a \in A$ is paired with a player choosing $b \in A$.

We now ask when a pure strategy $a^* \in A$ can be a stable equilibrium in the sense that it cannot be invaded by another strategy/phenotype $b \in A$. Suppose that a very small share of b mutations, $\epsilon > 0$ but very close to zero, appears in the population. The expected (or average) fitness of a^*-types is then greater than the expected fitness of the b-types if

$$(1 - \epsilon)u(a^*, a^*) + \epsilon u(a^*, b) > (1 - \epsilon)u(b, a^*) + \epsilon u(b, b).$$

If this condition holds the mutant will not be able to invade successfully.

Notice that the inequality will be satisfied for all $b \in A$, $b \neq a^*$, and sufficiently small ϵ if either

1. $u(a^*, a^*) > u(b, a^*)$ OR
2. $u(a^*, a^*) = u(b, a^*)$ and $u(a^*, b) > u(b, b)$

for all $b \neq a^*$.

The first condition is the usual Nash equilibrium condition, but with a strict inequality (or preference) rather than a weak one. In fact it is useful to distinguish what are known as *strict Nash equilibria*.

Definition 15. *For normal-form game* $\Gamma = \langle I, (A_i), (\succeq_i)\rangle$ *with* $A = \times_i A_i$, $a^* \in A$ *is* **a strict Nash equilibrium** *if for all i,* $(a_i^*, a_{-i}^*) \succ_i (a_i', a_{-i}^*)$ *for all* $a_i' \neq a_i^*$.

In the insect game above, (X, X) is a strict Nash equilibrium, while (Y, Y) is not. *Strict Nash equilibria are always evolutionarily stable* (cannot be invaded by a rare mutation).

The second condition is different, and stronger, than Nash stability, because it requires that a stable strategy do better against the mutation than the mutation does against itself, in addition to doing equally well against itself as the mutation does (which is the Nash condition with a weak best reply). (Y, Y) in the insect game does just as well against Y as against X, but worse against X than X against itself, so that even a tiny fraction of X's in the population would proliferate over time.

A strategy $a^* \in A$ that satisfies either of these two conditions is called an *evolutionarily stable strategy*, or ESS (Maynard Smith, 1982). ESS is a refinement of Nash equilibrium because every ESS is Nash but not all Nash equilibria are ESS.

Mixed strategies. You might wonder if it is possible for there to be an evolutionarily stable (and thus also Nash) equilibrium in which there is more than one phenotype/action in the population – a "polymorphism," in biology terms. The answer is Yes, and this is a good way to introduce the idea of a mixed strategy Nash equilibrium.

Let the available actions/phenotypes be $A = \{Hawk, Dove\}$, describing the genetic disposition of an animal to fight over territory or mates. Many animal species, including many bird species, are constantly deciding whether to contest others for territory or mates, which can gain a "prize" but also risks costly damage. Biologists have represented the basic situation with a simple 2×2 game that they call Hawk-Dove.[26]

Suppose that the fitness value or payoff for gaining the territory or mate is $v > 0$, and the expected fitness cost of fighting is $c \in (v/2, v)$. Suppose that if both are Hawk types, each has a half chance of winning the prize, and that if both are Dove types, both back off but whoever backs off first cedes the prize to the other, with ex ante equal chances. A Dove cedes the prize to a Hawk without any fight. So we have

		2	
		Dove	*Hawk*
1	*Dove*	$v/2, v/2$	$0, v$
	Hawk	$v, 0$	$v/2 - c, v/2 - c$

This strategic situation is clearly similar to Battle of the Sexes – a coordination game with two asymmetric pure-strategy Nash equilibria over which the players have conflicting preferences.[27] Without

[26] Known to Political Scientists by a different avian analogy, Chicken, referring to the deadly dare game played by James Dean and Corey Allen in *Rebel without a Cause*, and argued by many IR scholars to represent the strategic problem faced by state leaders in a nuclear crisis like the Cuban Missile Crisis. See also the strange 1980s version in *Footloose*.

[27] It is not the same as BoS. The strategic differences are subtle, though, and need not concern us here.

a means to distinguish each other, however, neither human nor non-human animals will be able to implement an efficient asymmetric equilibrium (as per discussion above).

Notice that a population composed entirely of Dove types would not be evolutionarily stable (or Nash), since a small number of Hawk-types could invade and get approximately $v > v/2$ in each match. But neither is a population composed completely of Hawk types stable, since a small number of Doves would have an expected fitness of approximately $0 > v/2 - c$.

Exercise 9. *Let $h \in [0, 1]$ represent the proportion of Hawk types in the population. Calculate the expected fitness of Hawk type and a Dove type, and compare. When would the share of Hawks in the population be increasing? When would the share of Doves be increasing? What does this imply? How does the stable distribution vary with the fitness parameters v and c?*

As you found, there is a unique share of Hawk types, h^*, such that the population distribution is stable. With a little more work you can check that this distribution in fact satisfies the second criterion for an ESS.

A critical feature of this stable distribution is that *expected fitness or payoff for each strategy that has positive probability is the same*. If one action had a strictly greater or smaller expected payoff, the shares would be subject to change over time in an evolutionary context, or would be chosen with probability 1 or zero in a "rational choice" context.

In this large population, matching-game interpretation, the shares of Hawks and Doves are interpreted as frequencies of actions/phenotypes that a player encounters in the world. Given h^*, each player gets the same payoff from Hawk and Dove and so either one is a best reply.

Exercise 10. *Carry out the same analysis as you did for Hawk-Dove for the following Stag Hunt game, where x is a parameter strictly between zero and 1. How many pure and mixed strategy Nash equilibria are there? Which are stable in the sense of ESS? If the situation started out with animals or players equally likely to choose C or D, and the number of players is very large, which equilibrium would be selected over time and how does this relate to x? Do you see any similarities to our analysis of the revolution problem in Section 9.7?*

		2	
		C	D
1	C	$1, 1$	$0, x$
	D	$x, 0$	x, x

12 Expected utility essentials

12.1 Cardinal versus ordinal utility functions

For a morning beverage Marilyn prefers coffee to tea and water, and tea to water: coffee $\succ$ tea $\succ$ water. Table 1 displays five utility functions that represent these preferences.

Table 1: Five utility functions on A = {coffee, tea, water}

	u	v	w	z	z'
coffee	10	10	1	15	-3
tea	8	3	.8	8	-4
water	0	0	0	5	-8

Understood as *ordinal* utility functions, all five encode exactly the same information about Marilyn's preferences. The specific numerical values are completely irrelevant except for how they order, or rank, the outcomes.

If I told you that these were *cardinal* or *expected utility* functions, I would be saying that they encode some additional information about Marilyn's preferences beyond her ranking of the three drinks. Consider u in Table 1. As a cardinal utility function, this says that Marilyn would be just indifferent between getting tea for sure, and a lottery or gamble that gave her coffee with probability .8 and water with probability .2. It says further that if p is the probability of getting coffee in a lottery on coffee and water, then Marilyn strictly prefers the lottery (p coffee, $(1 - p)$ water) to tea if $p > .8$, and strictly prefers tea to the lottery when $p < .8$.

If Marilyn's cardinal preferences are represented by v, then this means that she is just indifferent between getting tea for sure and the lottery that gives coffee with probability .3 and water with probability .7.

Substantively, u-Marilyn likes coffee better than tea in the morning, but not that much more relative to water. We could say that u-Marilyn's preference for coffee versus tea is not that intense, relative to water. Or we could equally well say that her dislike of water, relative to coffee or tea, is fairly intense (perhaps because it's about getting at least some caffeine).

v-Marilyn, by contrast, is willing to run a much greater risk of getting water in hopes of getting coffee, versus settling for tea. So we can say that v-Marilyn has a more intense preference for coffee relative to tea and water. Or we could equally well say that she doesn't have a much stronger preference for tea versus water relative to coffee.

Expected utility adds information about a person's intensity of preference over different outcomes, where intensity is understood in terms of preferences over different possible lotteries (or gambles) on the outcomes. Whereas an ordinal utility function represents preferences over outcomes $a \in A$, a cardinal utility function represents a person's preferences not only over A but also over lotteries or gambles on A, that is, over all probability distributions $\alpha \in \Delta(A)$.

Why are we doing this? Because it is a common and fundamental feature of strategic situations that a person is unsure about what other people are going to choose, and we want to be able characterize and represent their preferences over actions that can have different results depending on the uncertain actions that others take. Consider Figure 3a and suppose that player 1 has the belief that player 2 is

Figure 3: An EU function

1 \ 2	L	R
U	coffee	water
D	tea	tea

(a)

1 \ 2	L	R
U	10	0
D	8	8

(b)

going to choose L with probability .5. Then if 1's preferences on lotteries on {water, tea, coffee} are represented by u, she prefers the action D, and if her preferences are represented by v, then she prefers U.

Replacing the outcomes with the utility numbers in the case of u from Table 1 yields Figure 3b, which is convenient because now we can say what action(s) player 1 would prefer for any given belief about the likelihood that player 2 chooses L, if her preferences over lotteries on $A = \{\text{coffee, tea, water}\}$ are represented by u.

For instance, let $\alpha = (p, 0, 1-p)$ be the lottery on A that gives coffee with probability p, tea with probability zero, and water with probability $1-p$. The *expected utility* of U for player 1 is then $pu(\text{coffee}) + (1-p)u(\text{water}) = p10 + (1-p)0 = 10p$. This is greater or less than the expected utility of D (tea for sure) as $10p \gtrless 8$ or $p \gtrless .8$.

Next, observe that utility function w in Table 1 represents exactly the same cardinal preferences as utility function u, because it likewise implies that Marilyn prefers lottery $\alpha = (p, 0, 1-p)$ to "degenerate lottery" $\beta = (0, 1, 0)$ just as $p \gtrless .8$. You can work out that z represents the same EU preferences as v, and z' is equivalent in this sense to u. With cardinal utility what matters is both the order and the *relative spacing* between the utility numbers. The relative spacing encodes information about preferences over gambles and thus intensity of preference. z' is equivalent to u because -4 is 8/10's of the way between -8 and -3. z is equivalent to w because 8 is 3/10ths of the way between 5 and 15.

More generally, the set of all lotteries on outcomes A is $\Delta(A)$, with typical element $\alpha = (\alpha^1, \alpha^2, \ldots, \alpha^n)$. These are the probabilities of the lottery yielding each outcome $a^1, a^2, \ldots, a^n$ in A. We say that

Definition 16. *A person's preferences over lotteries on the set $\Delta(A)$ can be represented by an* **expected utility function** *$u : A \to \mathbb{R}$ if for all $\alpha, \beta \in \Delta(A)$, $\alpha \succeq \beta$ if and only if*

$$EU(\alpha) = \sum_{k=1}^{n} \alpha^k u(a^k) \geq EU(\beta) = \sum_{k=1}^{n} \beta^k u(a^k).$$

Exercise 11. *Suppose that a person's preferences $\succeq$ on $\Delta(A)$ can be represented by the expected utility function $u : A \to \mathbb{R}$. Show that her preferences can also be represented by any function $v(a) = cu(a) + d$ where $c > 0$.*

Above I used $EU(\alpha)$ for "the expected utility of lottery α." A very common and useful "abuse of notation" is to write $u(\alpha)$ for this. It is to be understood that this refers to the expected utility $\sum_k \alpha^k u(a^k)$,

where $u(a)$ is the utility number assigned to the sure outcome a.[28]

For our normal-form game models, we will use this same abuse, a lot. For instance, consider a two-player normal-form game in which player 1 has n pure strategies and player 2 has m pure strategies. For player 1's expected utility when 1 chooses $a_1 \in A_1$ and player 2 chooses the mixed strategy $\alpha_2 \in \Delta(A_2)$, we use

$$u_1(a_1, \alpha_2) = \sum_{l=1}^{m} \alpha_2^l u_1(a_1, a_2^l).$$

Likewise, for player 1's expected utility when 1 plays the mixed strategy $\alpha_1 \in \Delta(A_1)$ and 2 plays $\alpha_2 \in \Delta(A_2)$, we use

$$u_1(\alpha_1, \alpha_2) = \sum_{k=1}^{n} \alpha_1^k \sum_{l=1}^{m} \alpha_2^l u_1(a_1^k, a_2^l).$$

12.2 Expected utility versus expected value

Suppose I offer you a choice between \$5,000 and a lottery that gives you a p probability of \$10,000 and a $1 - p$ chance of zero. What probability p would make you just willing to take the gamble? For most people p is greater than one half.

This is so despite the fact that if you could play the gamble repeatedly, with a $p > 1/2$ your average payoff over many plays would be close to $\$10{,}000p > \$5{,}000$. $\$10{,}000p$ is the *expected value* of this gamble. But surely there is nothing necessarily irrational about preferring \$5,000 for sure to a p chance at \$10,000, if this is a one-shot choice and p is not much larger than .5.

As you may know from statistics, the expected value of a random variable X that takes (numerical) values $x_1, x_2, \ldots, x_n$ with probabilities $p_1, p_2, \ldots, p_n$ is $E(X) = \sum_i p_i x_i$. Like an expected utility, an expected value is a weighted average of numbers, with the weights being probabilities. An expected utility is in fact the expected value of some utility numbers.

Observe that a gamble such as (p coffee, $1 - p$ water) does not have an expected value because these outcomes have no natural numerical metric that is the relevant consideration for a person's preferences. This is often true of the outcomes we consider in political economy models. For example, although we may be able to range policies on abortion or gun control on some kind of dimension – from most restrictive to least restrictive, perhaps – policy descriptions will be (a) discrete and (b) lacking objective features that allow one to say things like "the difference between policy A and B is twice as large as the difference between policy B and C."

Sometimes outcomes of interest do have a natural metric, such as tax rates or any monetary outcome, like disposable income. And in fact much work in political economy imagines as an analytic convenience that a policy can be represented as a real number or a vector of real numbers, even when there really is no objective, measurable feature of the policy that the numbers would correspond to. For instance, below we will constantly work with "policy dimensions" represented as the $[0, 1]$ interval,

[28] Formally, we are saying that $u : \Delta(A) \to \mathbb{R}$, but allowing $u(a)$ to mean that the argument is the degenerate lottery that gives $a \in A$ with probability 1.

thus assuming a continuous metric over which one can define expected values. A great deal of work – and indeed a great deal of media and policy discourse – imagines a continuous Left-Right ideological spectrum representable by real numbers. There are even scaling methods that purport to estimate politicians' positions on an interval scale based on their roll-call voting records in the US Congress, or functions of the set of their campaign donors.

We saw from the example above that it can be normal and reasonable for a person to prefer getting the expected value of a money lottery to the lottery itself. There, most people would have \$5,000 $\succeq$ (.5 \$10,000, .5 \$0). When this is so we say the person is *risk averse*.

More generally, consider a set of numerical outcomes X that is an interval, say $X = [a, b]$, and consider lotteries on X represented by cumulative distribution functions $F(x)$. Let $E(F)$ refer to the expected value of the lottery F, which is $\int_a^b x dF(x)$. Suppose that person i has preferences over lotteries on X that can be represented by an expected utility function u_i, with the expected utility of lottery F defined as $u_i(F) = \int_a^b u_i(x) dF(x)$ (again abusing notation re u_i). Let δ_x be the *degenerate lottery* that gives x with probability 1, thus the "sure thing."[29]

Definition 17. *Person i is* **risk neutral** *with respect to X if she has $\delta_{E(F)} \sim_i F$ for all lotteries F. She is said to be strictly* **risk averse** *if $\delta_{E(F)} \succ_i F$ and strictly* **risk acceptant** *if $\delta_{E(F)} \prec_i F$ for all F.*

In words, a person's preferences exhibit risk aversion with respect to gambles on a numerical outcome if she prefers getting the expected value of every non-degenerate gamble to the gamble, just as you probably would prefer \$5,000 to the 50-50 gamble on \$10,000 or zero.

Suppose we were to extend this experiment so that I asked you the following: For each $x \in X = \{0, \$1k, \$2k, \ldots, \$9k, \$10k\}$, what is the smallest probability p_x such that you would just be willing to take the gamble $(p_x\ x, (1 - p_x)\ 0)$ in preference to getting a for sure? Figure 4 graphs hypothetical responses for a risk averse, risk neutral, and risk acceptant person. Observe how the risk averse person strictly prefers $\$x$ for sure to the lottery that gives her \$10k with probability x, whereas the risk-neutral person is indifferent and the risk acceptant person rejects the sure thing in preference for the gamble.

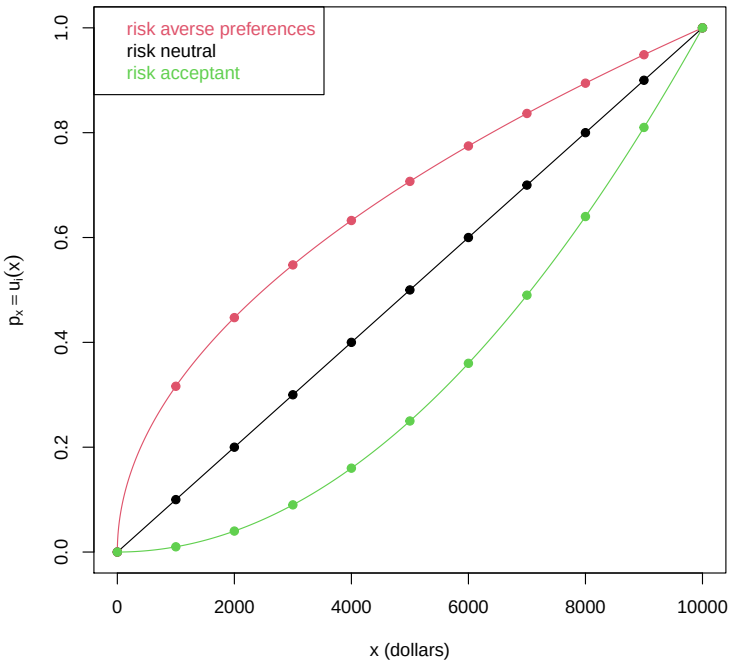
You will also notice that fleshed out as a continuous function $u_i(x)$, strictly risk-averse preferences appear strictly concave while risk-neutral preferences are linear. This is true in general: EU preferences $\succeq$ on outcomes $x \in [a, b]$ are risk averse (neutral, acceptant) if and only a $u(x)$ that represents them is concave (linear, convex).

12.3 Absolute versus relative risk aversion

In the International Relations literature, scholars often say things like "Hitler was risk acceptant about war," or "Powell was more risk averse [about going to war in Iraq] than Cheney." Because these statements refer to gambles over outcomes that do not have any clear numerical metric, it is not clear what to make of them using the concept of risk attitude just presented, which is sometimes known as *absolute risk attitude*.

[29] That is, F such that $F(z) = 0$ for $z < x$ and $F(z) = 1$ for $z \geq x$.

Figure 4: Risk attitudes for EU preferences defined on numerical outcomes



But it is perfectly sensible to interpret "Hitler was risk acceptant about war" as follows. By this the author means that in a series of international crises that he instigated in the 1930s, Hitler was willing to accept gambles with war as the bad outcome that his generals did not want to take. So he was *relatively* risk acceptant. Or, "Powell was more risk averse than Cheney" just means that the Cheney would have been willing to go to war in Iraq with a lower chance of a good outcome than Powell would have.

This idea of risk attitude is sometimes known as *relative risk aversion* (or acceptance).[30] The set of outcomes $A = \{$win the war, status quo (no war), lose the war$\}$ may not have any natural numerical metric, as in the case of coffee, tea, and water. But it makes sense to say that person A is relatively more risk averse than person B if both have win $\succ$ status quo $\succ$ lose but person A requires a higher chance of winning to prefer war to the status quo than person B does.

More generally, consider two people, 1 and 2, who share preferences $a_1 \succeq a_2 \succeq \ldots \succeq a_n$ over a set of n outcomes, with at least one of these also strict. We could say that 1 is relatively more risk acceptant than 2 if whenever $\alpha \succ_2 a_i$ for some $\alpha \in \Delta(A)$ and $a_i \in A$, it is also true that $\alpha \succ_1 a_i$, but there are also lotteries and outcomes such that $\alpha \succ_1 a_i$ but $a_i \succ_2 \alpha$.[31]

[30] Regrettably, there are also concepts of "absolute" and "relative risk aversion" defined for cases where the outcomes have a metric over which preferences apply. ...

[31] [I haven't seen this and I am not sure if it is right.]

12.4 Why EU?

Recall that players' preferences, not utility functions, are primitives in the normal-form model of a strategic situation, $\Gamma = \langle I, (A_i), (\succeq_i)\rangle$. Utility functions are for our convenience as analysts, not things that we assume are mentally calculated as such by actors, be they people, organizations, birds, or computer programs.

We saw earlier that a player's preferences $\succeq_i$ on a finite set of outcomes A can be represented by an ordinal utility function if and only if they are complete and transitive. We have now posited that a player has preferences $\succeq_i$ over all lotteries α in the set $\Delta(A)$, and proposed a numerical representation in the form of expected utility. If we do not imagine that people are introspectively deriving utility numbers and calculating expected utilities, then why think that people's preferences over lotteries would tend to be ordered in such a way that it is *as if* they were doing this?

John von Neumann and Oskar Morgenstern (1947) proved that if a person's preferences over lotteries in $\Delta(A)$ accord with two additional consistency conditions (axioms) beyond completeness and transitivity, then they can be represented by an EU function. These axioms put conditions on a person's preferences between *compound lotteries*, which are lotteries on lotteries. For example, if $\alpha, \beta \in \Delta(A)$ are two lotteries, then $(p\alpha, (1-p)\beta)$ is the compound lottery that gives α with probability p and β with probability $1-p$.

Definition 18. *$\succeq$ on $\Delta(A)$ satisfy axiom* **A3** *("Archimedean", or "continuity") if for any lotteries $\alpha, \beta, \gamma \in \Delta(A)$ such that $\alpha \succ \beta \succ \gamma$, there exist numbers $p, q \in (0,1)$ such that $(p\alpha, (1-p)\gamma) \succ \beta \succ (q\alpha, (1-q)\gamma)$.*

They satisfy **A4** *("substitution," or "independence") if for all $p \in (0,1)$, $\alpha \succ \beta$ implies that $(p\alpha, (1-p)\gamma) \succ (p\beta, (1-p)\gamma)$ for all $\gamma \in \Delta(A)$.*

Substantively, your preferences satisfy the continuity axiom if there is no outcome or lottery γ that you see as so bad that if you prefer lottery α to β, then you would reject any and all compound lotteries on α and γ in preference to β. And likewise there is no outcome or lottery α that you like so much that you prefer any compound lottery in which α gets positive probability to β. It is sometimes referred to as the "no heaven, no hell" axiom. It does not seem very demanding by itself, but it does turn out to imply that if a person's preferences respect it, then there is a unique probability p such that $(p\alpha, (1-p)\gamma) \sim \beta$ for any α, β, γ such that $\alpha \succ \beta \succ \gamma$. This allows "calibration" of an expected utility function in von Neumann's proof.

The independence axiom says that if you like lottery α better than β, then this is still true if we add to both options a positive chance (less than 1) of getting some other lottery or outcome. Figure 5 illustrates. The axiom says that if you have $\alpha \succ \beta$ then for any probability $p > 0$, you must also prefer U to D in this choice problem. The idea is that U and D here differ only with respect to α and β, so the fact that there is a $1-p$ chance of ending up with γ in both option U and option D should not change your evaluation of α versus β. This also seems pretty mild and reasonable, and probably holds most of the time for most people facing relatively simple choice problems (as in the figure).

But people are not axiomatic decision-makers in the sense of mechanically and rigorously applying general, abstract rules when choosing among risky and uncertain options. Instead, we develop heuris-

Figure 5: The independence axiom: $\alpha \succ \beta \Rightarrow U \succ D$ for any $p \in (0,1)$

		p	$1-p$
1	U	α	γ
	D	β	γ

tics for real-world problems that we encounter frequently, and adapt these heuristics on the fly, using analogies, when we encounter novel situations. There is a large literature on how people tend to deviate from the strict implications of the EU axioms in certain classes of choice problems. This originated with work by Kahneman et al. (1982) in Psychology and has blossomed into much research in "behavioral economics" that explores implications of common heuristics and biases for choice-making in political-economy situations.

In my view EU remains extremely useful for analyzing and understanding strategic situations in political economy, at the least as a point of departure. This is so for two main reasons. First, the EU representation captures, in a simple and highly tractable form, the idea that a gamble is more attractive for a person the more probability weight it puts on relatively good outcomes and the less weight it puts on relatively bad outcomes. This principle is overwhelmingly plausible. EU implies a good deal more than this, so it should not be taken too literally in terms of making point predictions. But for "more of this implies more (or less) of that"-type comparative statics, this is quite often going to be enough.

Second, there are reasonable arguments to the effect that if a person's preferences on $\Delta(A)$ violate the EU axioms, this indicates irrationality – the person would want to change their expressed preferences if the violations were pointed out and the implications explained. So when we are asking the *normative* question "How should a rational person choose in such-and-such strategic situation?" there are some grounds for assuming EU-consistent preferences.[32]

13 Bargaining problems: Schelling vs Nash

Politics obviously involves a great deal of bargaining in the ordinary-language sense of negotiations in which people make offers or demands on how to coordinate or divide up a resource. Some examples:

- In US legislative politics, party leaders constantly bargain with their own members and with leaders from the other party over the terms of bills. Failure to agree on terms may leave a status quo in place that is disliked by a majority, or in some cases by almost everyone. In the case of annual budget resolutions, failure to agree means a government shutdown, which can be risky and costly for both parties in various ways, including voter animosity. Nonetheless, government shutdowns occur, as the two sides hope that the costs and risks of shutdown will cause the other to make concessions. This kind of costly delay or refusal as a pressure tactic occurs often in bargaining over other types of legislation as well.
- Within parties, most legislators may want to move current policy to the left (Democrats) or the

[32] [references to money pump arguments]

right (Republicans). But they always have conflicting preferences over how far to move. Moderates prefer less change than "extremists." When an election leads to a change in House majority party, there is a lot of bargaining within the party over where to position the new majority's legislative agenda.[33]

- There is a natural sense in which domestic politics – "politics" – is always one big bargaining problem: People have conflicting views on many different policy dimensions, but jointly (or in large majorities) prefer a range of compromises to a bloody civil war, or less costly but still mutually hurting outcomes.

- In international politics, state leaders bargain with each other over the terms of trade deals, foreign direct investment, tax treaties, extradition treaties, border placement, environmental agreements, human rights conventions, arms control, the design and policies of international organizations, and much more. Lack of agreement has different consequences in different issue areas and settings. But, in general, failure to agree – or delay in agreement – entails economic or political costs for the parties to the interaction. In interstate and civil wars, combatants suffer extreme costs in efforts to convince the other side to make concessions of various sorts (or defeat them outright).

One would hope that game-theoretic analysis would have something interesting to say about bargaining, since it is a clearly a strategic situation in the sense of section 6, and because it is so ubiquitous and important in human life (not just politics). Game theory does not disappoint, although some of the main conclusions imply pessimism about the possibility of predicting bargaining outcomes in the absence of specific contextual detail like that discussed in section 10 on correlated equilibrium.

Our everyday concept of bargaining imagines a dynamic process, not a static, simultaneous normal-form game. We imagine two people making and responding to offers through time. I make you an offer, you respond with No, how about this, and we continue till someone says Yes. Bargaining also frequently involves threats like "this is my final offer, take it or leave it." That is, bargainers may try to commit themselves to not budge from a position, to carry out other costly actions if the other side does not concede, or to maneuver to put the other side in the position of having "the last clear chance" to avoid no deal (Schelling, 1960, chapter 2).

Extensive-form games are needed to model bargaining as a dynamic process and to investigate questions about credibility of threats and promises in bargaining. We consider such models in Part 5.

But even within the constraints of the normal form, we can get some important insights. The insights I will cover here derive from work on the bargaining problem in late 1940s and the 1950s by John Nash and Thomas Schelling. To summarize:

- Schelling saw the essence of bargaining in the fact of multiple Nash equilibria, so that the problem from the individual bargainer's perspective was predicting how to coordinate. That is, he sees bargaining problems as fundamentally problems of *equilibrium selection*. He did not think that there was any strong reason to expect that bargainers would necessarily converge on common expectations of a particular equilibrium in bargaining problems. But he did think that people

[33]The House Republican's saga after the 2022 election gave them a nominal majority is a dramatic case in point.

could be surprisingly good at using contextual features and commonly known principles – "focal points" and (what might be called) focal principles – to solve the problem of equilibrium selection in bargaining situations.

An important implication of Schelling's view is that trying to prescribe behavior or make predictions about bargaining outcomes based only on utility functions (individuals' preferences) is *a fool's errand*.

- Nash (1953) in effect took up the fool's errand, proposing an equilibrium selection argument for the bargaining problem that contains the following insight: The player who is more willing to risk disagreement (coordination failure) gets more. Willingness to risk disagreement in Nash's approach is based entirely on preferences over divisions of the pie, as represented by the players' expected utility functions. The argument is more of a positive prediction than a normative claim about how "pies" should be divided.

I think that Schelling and Nash are both right. This is a case of "both/and" rather than "either/or." Nash is correct that in most if not all bargaining situations, parties can if they wish take actions that generate risk of disagreement, or what Schelling (of all people) called "threats that leave something to chance." They can also take actions that delay agreement, which can be costly. Consistent with everyday experience and myriad examples from politics, greater willingness to run risks or pay costs does regularly translate to getting a better deal.

But Schelling is also right that how to coordinate on one of many possible resolutions is the strategic heart of bargaining problems, and that specific contextual features and focal principles outside the spare game description very often determine equilibrium selection and thus the outcome. Think, for example, of how often a 50-50, split-the-difference "logic" prevails. Or how society-specific rules of deference and courtesy shape bill-splitting or the allocation of decision-making authorities in households. Or how often a precedent – how it was done last time – governs the outcome.

Curiously, Schelling has almost nothing to say about the moral or ethical dimensions of focal points and focal principles, even though "50-50" and more contextual, conditional norms of fairness are not just rules of thumb, conveniences for avoiding coordination failures. In many contexts people endow them with great moral or ethical significance, arguing over what is the *right* or *just* division of the pie in their appeal to focal moral principles.

In turn, convictions about what just division is in particular contexts affect the parties' willingness to run risks of disagreement, for reasons separate from the raw preferences over amounts of pie. Put differently, one could say that highly contextual moral considerations and arguments determine the shape of one's "utility function for money" (for example), such that $u(x)$ is not the same at all across contexts.

I have not seen this kind of consideration fully formalized. It is related to what psychologists and behavioral economists call "framing effects," although those are usually conceived of as hard-wired cognitive patterns or errors. One could write $u_i(x, c)$, where x is an amount of (say) money, and $c \in C$ is a "frame" or "moral context." The idea is that a person's value for what she gets in the division of a resource in a bargaining situation can depend not only on amount of the resource itself, but also on her views about what is right, or just, or appropriate in this particular situation. This is different from

Schelling's idea that context works as a coordinating device, as in correlated equilibrium.

13.1 Political bargaining, or bargaining in the shadow of power

The paradigmatic example of bargaining in economics occurs between a buyer and a seller. For example, the buyer values the object at $\bar{v} > \underline{v} \geq 0$, where $\underline{v}$ is the seller's value for not selling the good. Any price $p \in (\underline{v}, \bar{v})$ makes both parties better off than no deal, which has payoffs set to zero for the buyer and $\underline{v}$ for the seller. The buyer and seller have conflicting preferences over p, since payoffs for a deal are $(\bar{v} - p, p - \underline{v})$. But they have a common interest in striking some deal.

One of the most common mistakes, or confusions, that political scientists and many others display when talking about politics is to describe various bargaining situations as "zero sum." It is true that, for instance, $\bar{v} - p + p - \underline{v}$ is a constant, so that a buyer's and seller's preferences over the price are strictly competitive and thus in a sense "zero sum." But the strategic situation is emphatically *not* a zero- (or constant-) sum problem, because the parties have a common interest in coordinating on some agreement rather than no agreement.

In buyer-seller bargaining, the natural assumption is that disagreement means that the parties "walk away." There is no trade. This can be the natural assumption in many political contexts as well, as when Democratic and Republican party leaders bargain over the terms of the bill which, if not passed, will leave a defined status quo in place.

But in political contexts it often happens that "no deal" implies, or can imply, a costly fight in which success means that the winner can unilaterally determine the outcome. For example:

- The leaders of states bargain over territory and many other things with the use of force in the background in case of disagreement.
- During wars, states (and non-state armed groups, in civil conflicts) endure the costs and risks of fighting while bargaining over terms of a settlement.
- In US politics, presidents and party leaders in Congress bargain over budgets or authorizations to increase the debt ceiling in shadow of a government shutdown, or increased default risk and financial sector catastrophe. A government shutdown will have, to some degree, hard-to-predict effects on public opinion and who gets the lion's share of the blame. It thus has an aspect of a costly lottery.
- In parliamentary systems, the possibility of calling an election can act as a disagreement point for bargaining between or within parties in the legislature.

Let's examine a model of interstate bargaining in which we suppose that failure to reach agreement implies a costly military conflict in which state 1 defeates state 2 with probability $p \in [0, 1]$. Thus p represents state 1's relative military capabilities, or power.

States 1 and 2 have conflicting preferences over the resolution of some issue, perhaps the control of territory, that we represent with the interval $X = [0, 1]$. In a simple and risk-neutral case, state 1's

value for outcome $x \in [0,1]$ is x and state 2's is $1-x$. More generally we could have 1's preferences represented by $u_1(x)$ and 2's by $u_2(1-x)$, where these utility functions are strictly increasing in their arguments and are normalized so that $u_i(0) = 0$ and $u_i(1) = 1$.

Assuming that the winner of a war chooses the issue or territorial resolution as it pleases, expected utilities for a fight are then $pu_1(1) + (1-p)u_1(0) - c_1 = p - c_1$ for state 1 and, similarly, $1 - p - c_2$ for state 2. c_1 and c_2 are (utility) costs of fighting.

Exercise 12. *Steps to more fully understand this set up:*

1. *Explain why, by expected utility theory, there is no harm – or "loss of generality" – in normalizing the utility functions u_1 and u_2 so that $u_i(0) = 0$ and $u_i(1) = 1$, for $i = 1, 2$.*
2. *But what implication does this normalization have for the interpretation of c_i? Hint: It may help to think about a variation in which state i's utility function is represented as $w_i(z) = v_i u_i(z) - c_i$, where $v_i > 0$ parameterizes "value for territory relative to costs of fighting."*
3. *Identify, for the risk-neutral case $u_i(z) = z$, the set of deals $x \in [0,1]$ that both prefer to a fight.*
4. *Do the same for concave (risk averse) utility functions $u_i(z)$. Notes: Concavity implies that $u_i(z) > z$ for $z \in (0,1)$. You will need to use the inverse functions $u_i^{-1}(u_i)$ to express your answer.*
5. *For convex (risk acceptant) utility functions (which implies $u_i(z) < z$ for $z \in (0,1)$), identify the set of lotteries on $x \in \{0,1\}$ that both prefer to a fight.*
6. *For the risk-neutral case, graph u_2 as a function of u_1 for $x \in [0,1]$, adding the disagreement point $(p - c_1, 1 - p - c_2)$. (This graph has u_2 on the y axis and u_1 on the x axis.) Where do you see the set of agreements both prefer to a fight, and how does this set change as you vary relative power p, and costs c_1 and c_2?*
7. *Consider the specific risk-averse functional form $u_i(z) = z^{\alpha_i}$ where $\alpha_i \in (0,1]$. Repeat the last question. Begin by expressing u_2 as a function of u_1. Graph the cases of (a) $\alpha_1 = \alpha_2 = 1/2$, and (b) $\alpha_1 = 1$, $\alpha_2 = 1/2$.*
8. *Show that the Pareto frontier – the set of outcomes $(u_1(x), u_2(1-x))$ – when expressed as $u_2(u_1)$ is decreasing and strictly concave in u_1 (assuming $u_1(z)$ and $u_2(z)$ are strictly concave).*

You will have found that in the risk neutral case, the set of deals both weakly prefer to a fight is $x \in [p - c_1, p + c_2]$.[34] Thus, relative power (or prospects) in a fight "matters" in that greater fighting strength increases the minimum one requires to prefer a deal to costly conflict. Similarly, lower c_i, which in words means greater "resolve," or perception of the "stakes," increases a player's minimum acceptable deal.

This all makes good sense. It squares with the observation that status quo resolutions of a large range of issues in both international and domestic politics tend to favor those with more power or motivation in the event of costly conflict.

[34] If you were being careful, your answer would account for the boundaries of $[0,1]$ as well.

But given a "bargaining range" like $[p - c_1, p + c_2]$ – or $[\underline{v}, \bar{v}]$ in a buyer-seller case – what determines who gets how much *within the range*? This is "the bargaining problem" as conceived by Schelling and Nash. How will the surplus, which is the total value or utility above the value of disagreement, be divided?[35]

The next section presents Nash's answer. First, here is a useful distinction for understanding how bargaining theory applies to real-world examples.

1. Sometimes, disagreement, or no deal, is the status quo in the sense that to reap the benefits of a deal, the parties have to agree to take actions that they are not currently taking. Buyer-seller bargaining is the standard example. The parties in effect start in a situation of disagreement or no deal, and bargain over how to move to the Pareto-frontier in the graphs you drew in the above exercise. Many political bargaining situations are the same – for instance, can we make a deal on new legislature to make us both (or all) better off that the status quo?

2. In other situations, there is already a status quo division or agreement in place – for example, $q \in (p-c_1, p+c_2)$ – and "bargaining" is about threats or actions to coerce a change by deliberately imposing a mutually bad disagreement point. "Cede this chunk of territory or I may invade." Or, "I have started a war, which is costly and risky for both of us; you can stop it by ceding my demands." Or, "I am more willing than you to risk a US government default in this period when we don't know exactly how long the Treasury can or will manage to keep paying bills. So give me legislative concessions now."

   Once a party has taken actions to impose or revert to a mutually bad disagreement point, the bargainers are in situation (1), the classic bargaining problem. Before this, we observe "coercive bargaining" in the sense of threats to create costly disagreement, or steps into costly disagreement. This kind of situation is best analyzed with game-theoretic tools for studying issues of credibility and commitment, as in Part 5. But a strategic tradeoff and logic that is central to dynamic bargaining situations can be nicely captured in a normal-form simplification.

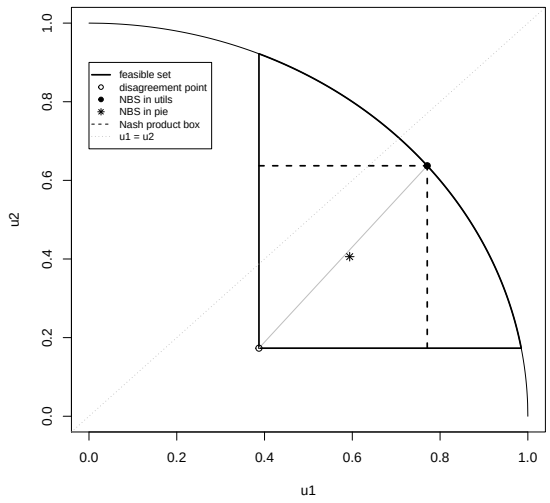
13.2 The smoothed Nash demand game

Recall the Nash demand game of Exercise 4: Two players simultaneously write down a demand, which is the share of the "pie," or resource, over which they are bargaining. If the shares sum to more than 1, then there is no deal. Otherwise, they get their demands. You observed that any pair of demands that sums to 1 is a Nash equilibrium, as well as there being an inefficient equilibrium which both demand the whole pie and there is no deal.

Nash (1953) showed that if we introduce a small amount of uncertainty into whether two demands will be compatible – for example, by making the size of the total pie available slightly uncertain – multiple equilibria go away and there remains, very generally, a unique pair of Nash equilibrium demands. These demands coincide with the Nash Bargaining Solution, another result due to a very different kind of approach (axiomatic) that Nash had also developed (Nash, 1950*a*).

[35] Surplus $= u_1(x) + u_2(1-x) - (d_1 + d_2)$, where d_i is i's disagreement utility. So, for example, $x + 1 - x - (p - c_1 + 1 - p - c_2) = c_1 + c_2$ in the risk-neutral political bargaining case.

Figure 6: The feasible set and disagreement point with risk-averse bargainers



The uniqueness result in the "smoothed" Nash demand game is almost always presented without any explanation of the strategic logic behind it, as something of a miracle whose miraculous-ness is implicitly supposed to justify the Nash Bargaining Solution. But in fact it captures a strategic logic common to bargaining situations in the real world in an enormous range of settings.

We will consider a somewhat more general version of the Nash demand game of Exercise 4. Again, players 1 and 2 simultaneously write down demands $x_i \in [0, 1]$, $i = 1, 2$, which are proposals for the amount of some resource that each gets if the demands are compatible, meaning that $x_1 + x_2 \leq 1$. The total amount of the resources is normalized to 1.

Preferences will be represented by expected utility functions $u_i(x_i)$ which are increasing in x_i. In addition, each has a *reservation value* in the event of no agreement, which is $u_1(a)$ for player 1 and $u_2(b)$ for player 2. a and b are in units of the resource at issue, e.g., money. We assume that $a + b < 1$, which implies that there is a set of divisions and associated utility pairs $\{(u_1(x_1), u_2(x_2)) : (u_1(x_1), u_2(x_2)) > (u_1(a), u_2(b))\}$ that both prefer to disagreement. This set of utility pairs is called *the feasible set*.

Figure 6 illustrates for a case of risk-averse preferences represented by $u_i(x_i) = x_i^{\alpha_i}$, where here $\alpha_1 = \alpha_2 \in (0, 1)$.[36] Observe that we are graphing in "utility space" where points are (u_1, u_2), not the (x_1, x_2) space. Nash called $(u_1(a), u_2(b))$ *the disagreement point*.

As in the exercise, this Nash demand game has an infinite number of efficient Nash equilibria: Any pair $(x_1, x_2) \geq (a, b)$ such that $x_1 + x_2 = 1$, corresponding to any point on Pareto frontier in Figure 6. As Schelling stressed, bargaining problems are problems of coordination. What will determine which if any of these equilibria is selected, or focused on, by the players? Schelling argued that in the real world,

[36]The frontier is defined by $u_2 = (1 - u_1^{1/\alpha_1})^{\alpha_2}$. If the players are both risk-*acceptant* in amounts of the resource ($\alpha_i > 1$ here), the curve bows in. In consequence, both players prefer lotteries on $(1 - b, b)$ and $(a, 1 - a)$ to actually dividing things up. This makes the frontier, or set of Pareto-efficient expected payoffs from feasible lotteries the 45 degree line connecting these two points.

specific contextual features such as norms are absolutely critical for player's bargaining strategies and for bargaining outcomes, because they influence players' expectations and what is likely to be mutually expected.

By contrast, working in the program of "let's strip strategic situations down to their barebones," Nash and others proposed arguments about what should be expected to happen – and, in another formulation, what ought to happen – in a generic bargaining problem based *solely* on $u_1(x_1)$, $u_2(x_2)$, and the disagreement point. Thus no specific context, just players' preferences over the outcomes.

It is debatable whether this approach makes sense, in light of Schelling's observations and argument. But Nash's approach does identify one quite compelling and surprisingly general idea about what determines "who gets how much?" in bargaining situations. The strategic intuition behind the "Nash bargaining solution" is that bargainers will divide up the pie in shares that reflect *their relative willingness to risk disagreement*. The party more willing to risk disagreement, as encoded in their expected utility preferences $u_i(x_i)$, gets more.

Nash arrived at this result in two main ways. In my view the more interesting and compelling one was his demonstration that by adding an arbitrarily small amount of noise to the Nash demand game, the set of efficient pure-strategy equilibria shrinks from "any efficient division" to exactly one, unique Nash equilibrium in the limit! This unique equilibrium is the pair (x_1, x_2) that maximizes the product $(u_1(x_1) - u_1(a))(u_2(x_2) - u_2(b))$, which is known as *the Nash bargaining solution*.

The "smoothed Nash demand game" is exactly as above, except that the players choose their demands before the size of the available pie is realized. The size of the pie will be $1 + \epsilon$, where ϵ is a mean-zero random variable with cdf F.[37]

This introduces some risk that any given pair of demands will be incompatible. Further, making a bigger demand increases the risk – there is a risk-return tradeoff. Although there is no particularly natural story for why we would imagine the size of the pie to be uncertain, we can think of the formulation as a concise way of capturing that there are many factors in specific settings that create some risk of failure to reach an agreement, with higher risk associated with more aggressive demands.

In fact, there is nearby interpretation that makes a lot more substantive sense. Instead of the size of pie being variable and uncertain, we can think of the bargainers having imperfect control over the compatibility, and thus results, of their demands.

This interpretation actually squares well with Schelling's understanding of bargaining in the context he focused on the most. Schelling characterized bargaining in international disputes between nuclear-armed states as "competitions in risk-taking" in which all manner of factors, psychological and otherwise, created a risk of nuclear escalation, which bargainers cannot avoid while bargaining in a nuclear crisis. A more realistic motivation than uncertainty about the size of the pie would be that x_1 and x_2 represent not specific amounts, but rather how tough or aggressive a player is in the unmodelled, context-dependent bargaining process. Let a player's "effective demand" be $z_i = x_i + \epsilon_i$, where ϵ_i is a mean-zero random variable with a differentiable cdf F. The idea is that all kinds of context-specific things can happen that make one's demands in bargaining effectively "threats that leave something to chance" to some degree, whether one wishes this or not. Now, there is a probability $Pr(\epsilon_1 + \epsilon_2 > z_1 + z_2 - 1)$ of no

[37] Nash's and other treatments I have seen study a more general class of "perturbed" demand games.

deal, which effectively the same as Nash's version.

Returning to Nash's variant where the size of the pie is a random variable: Player 1's expected payoff for making the demand x_1 when the other demands x_2 is then

$$u_1(x_1, x_2) = (1 - F(x_1 + x_2 - 1))u(x_1) + F(x_1 + x_2 - 1)u_1(a)$$

which has the first-order condition

$$\frac{f(x_1 + x_2 - 1)}{1 - F(x_1 + x_2 - 1)} = \frac{u_1'(x_1)}{u_1(x_1) - u_1(a)}.$$

Likewise, player 2's first-order condition is

$$\frac{f(x_1 + x_2 - 1)}{1 - F(x_1 + x_2 - 1)} = \frac{u_2'(x_2)}{u_2(x_2) - u_2(b)}.$$

from which it follows that at an equilibrium where both are satisfied,

$$\frac{u_1'(x_1)}{u_2'(x_2)} = \frac{u_1(x_1) - u_1(a)}{u_2(x_2) - u_2(b)}.$$

To get an initial intuition for what this says, consider the risk neutral case of $u_i(x_i) = x_i$. Then this condition becomes

$$x_1^* - a = x_2^* - b,$$

which means that the players equalize the (approximate) surplus above their reservation values.

What happens as the amount of noise ϵ gets small in the sense that its variance approaches zero? Fix a small $\delta > 0$, and consider demands by player 1 such that $x_1 \leq 1 - x_2 - \delta$. The probability that such a demand is incompatible is at most $Pr(\epsilon < -\delta)$, which goes to zero as the variance of the noise ϵ goes to zero. So player 1's best reply is certainly at least $1 - x_2 - \delta$. Now consider demands $x_1 \geq 1 - x_2 + \delta$. As F collapses towards a "spike" at zero, these demands will be incompatible with probability approaching 1, so player 1 would certainly want to choose $x_1 < 1 - x_2 + \delta$. Since this argument holds for arbitrarily small δ and in the same way for player 2, in the limit players' best replies must approach $x_1^* + x_2^* = 1$.

Putting this together with the equilibrium conditions above, we have that in limit,

$$\frac{u_1'(x_1^*)}{u_2'(1 - x_1^*)} = \frac{u_1(x_1^*) - u_1(a)}{u_2(1 - x_1^*) - u_2(b)}. \tag{1}$$

Remarkably, this is exactly the first-order condition that defines the solution to

$$\max_{x_1} \; (u_1(x_1) - u_1(a))\,(u_2(1 - x_1) - u_2(b))\,,$$

the efficient division that maximizes the Nash product!

Let's illustrate again with the relatively clean risk-neutral case. Here, the combination of $x_1^* - a = x_2^* - b$

and $x_1^* + x_2^* = 1$ implies

$$x_1^* = \frac{1}{2} + \frac{a-b}{2} \text{ and } x_2^* = \frac{1}{2} + \frac{b-a}{2}.$$

So in the unique limit equilibrium of the smoothed Nash demand game with risk-neutral preferences, the players divide the surplus above their reservation utilities equally. And increasing the value of a player's "outside option" – the minimum they need to be willing to agree – leads to getting a larger share of the pie, even though equilibrium division gives the player strictly more than the outside option.

What is going on here? Nash's result shows that introducing an arbitrarily small amount of noise takes us from a coordination game with an *infinity* of equilibria to a game with a *unique* equilibrium that, in the case of $u_1(x) = u_2(x) = x$, predicts equal division of the surplus above reservation values. This should seem remarkable, even a bit miraculous. But there is a quite comprehensible strategic logic and intuition that drives it.

For convenience let $a = b = 0$. Let's think about why an unequal division, say $(1/4, 3/4)$, can be a Nash equilibrium when there is no noise, but is eliminated in favor of equal division when a little noise is introduced. With the noise, the players face the same risk of incompatible demands for any given pair of demands (x_1, x_2). Consider each player's willingness to risk "breakdown" (incompatible demands) in order to get an additional increment dx from the other player. At $(1/4, 3/4)$, player 1 is looking at a lottery on $1/4 + dx$ and zero, whereas player 2's lottery is on $3/4 + dx$ and zero. Because she has less to lose, player 1 *is more willing to risk breakdown by edging up her demand*. This would make player 2 too uncomfortable given how much he has to lose by disagreement (about 3/4). In other words, the player with less to lose relative to their reservation value is comfortable with a greater risk of breakdown, which means the other player prefers a lower demand to get a lower level of risk.

Thus, given the risk-inducing noise, a division (x_1, x_2) is going to be an equilibrium only when the players are *equally* willing to risk disagreement by pushing a little harder. In the general case of $u_i(x_i)$, this is just what equation (1) says. In words, their willingness to risk disagreement is balanced when the ratios of their marginal gain from pushing a little harder to their opportunity costs of breakdown are the same. That is,

$$\frac{u_1(x_1 + dx) - u_1(x_1)}{u_1(x_1) - u_1(a)} = \frac{u_2(1 - x_1 + dx) - u_2(1 - x_1)}{u_2(1 - x_1) - u_2(b)},$$

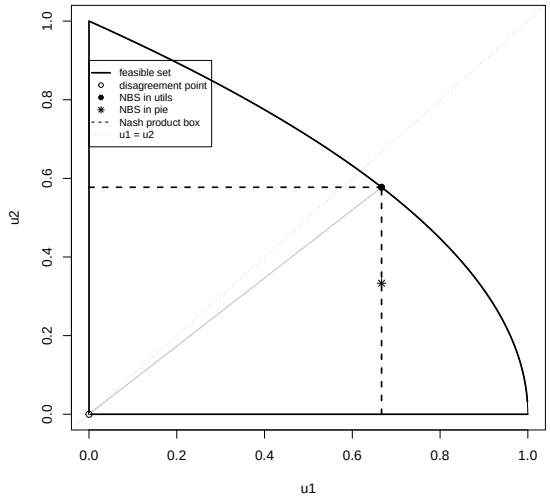
which is the same as (1) in the limit (after dividing both numerators by dx).

On the (u_1, u_2) graph, the Nash product is the area of the rectangle with corners at $(u_1(a), u_2(b))$ and $(u_1(x_1), u_2(1 - x_1))$ and sides parallel to the axes; see Figure 6. The unique limit equilibrium of the smoothed bargaining game occurs at the $(x_1, 1 - x_1)$ that maximizes the area of this rectangle. In graphical terms, the ratio $u_1'(x_1)/u_2'(1 - x_1)$ is the negative of the slope of the tangent to the Pareto frontier at $(x_1, 1 - x_1)$. The ratio on the right-hand side of (1) is the ratio of the sides of the Nash product box indicated in the figure. So the Nash bargaining solution is where the tangent to the frontier has the same slope as the diagonal line connecting the upper-left and lower-right corners of the box.

The substantive point, to reiterate, is that this is where the two players willingness' to risk disagreement is equalized.

Suppose that player 1 is risk neutral, with $u_1(x_1) = x_1$, while player 2 has concave, risk averse pref-

Figure 7: The Nash bargaining solution with risk-neutral player 1 and risk-averse player 2



erences $u_2(x_2) = \sqrt{x_2}$. Figure 7 shows what happens for the case of both reservation values at zero ($a = b = 0$).

We see that the relatively more risk-acceptant player, player 1, gets more pie. The shares are 2/3 for player 1 and 1/3 for player 2 in this case.

Exercise 13. *Find the Nash bargaining solution for $u_1(x_1) = x_1$ and $u_2(x_2) = x_2^{\alpha_2}$, $\alpha_2 \in (0, 1)$. Interpret comparative statics on α_2.*

13.3 How *should* the pie be divided?

The argument that leads to the Nash bargaining solution (NBS) in the smoothed Nash demand game is positive, not normative. The NBS outcome results from a balance of willingness to run risk of disagreement, rather than an argument or claims about what would be a fair or right division of the pie.

So let's ask next about how the pie *should* be divided. Notice first that in the basic Nash demand game, all of the efficient equilibria are Pareto efficient. So as a moral principle, Pareto is of no help for picking out any one division as socially best.

If you are wondering "How could I say what is "socially best" without knowing the specific social, political, or economic context of who is bargaining with whom over what?", then I would say that you are onto something. In the real world, our moral judgements are highly context-specific, whereas all we have in this stripped-down, schematic bargaining problem are individual preferences over amounts of the good or policy issue.

If you think about it, the idea that we could arrive at a theory of what is the fair or right way to divide the pie in the absence of specific information about the human context is completely absurd. I will

come back to this point, but after going through a famous, valiant, and instructive attempt to pretend otherwise.

The classical *utilitarian* idea is to associate "socially best" with the outcome that maximizes a sum of individual "utilities." In the case at hand, a naive utilitarian answer to the question of how the pie should be divided in the basic bargaining problem would be the x that maximizes $u_1(x) + u_2(1 - x)$. With risk-averse (concave) utility functions, this is the division x^* that solves the first-order condition, which in words is the division at which the players' marginal increases in utility from a small increase in "pie" are equal:

$$u_1'(x) = u_2'(1 - x).$$

But wait a minute, I hope you are saying. What does it mean to *add* two expected utility functions together? Precisely nothing. EU is a representation of a person's preferences over a set of outcomes. It says nothing about how one person's value for these outcomes compares to another person's. In the jargon, we have not yet made any assumptions that would justify or render meaningful *interpersonal comparisons of utility* numbers.

To illustrate why adding up EU numbers isn't meaningful on EU-theory terms, recall that if $u_i(x)$ represents person i's EU preferences, then so does any $v_i(x) = au_i(x) + b$ where $a > 0$ and b is any number. But the division of the pie that maximizes the utilitarian sum $v_1(x) + u_2(1 - x)$ is the x that solves $au_1'(x) = u_2'(1 - x)$.[38] This is obviously going to be different from x^* for any $a \neq 1$. In particular, the bigger a, the more weight we would effectively be putting on 1's preferences.

For a sum like $u_1(x) + u_2(1 - x)$ to be meaningful, we are going to have to say and assume more about what these functions represent – more than just a person's preferences over different lotteries.

The classical utilitarians did this by positing a thing called "utility" that is something like a level of happiness or wellbeing of an individual, and for which $u_i(x)$ is an index, or measure. For clarity, and only in this section, let's call this thing Utility, in contrast to "utility" in the sense of a convenient numerical representation of a person's preferences over lotteries on a set of outcomes.

Suppose that somehow we are able to assume, or that we are just willing to stipulate, that player 1's Utility for getting the whole pie ($x = 1$) versus getting none of it ($x = 0$) gives player 1 $a > 0$ times the Utility that player 2 would get from getting the whole pie, versus getting none of it. In the case of $a = 1$, we are saying that they get equal Utility (value, sense of satisfaction) for the getting the whole pie versus none of it.

If we are being Utilitarians, we might be willing to make this assumption because we think that as an empirical matter it is roughly correct, or that it is close enough given the difficulty of accurately revealing and measuring the true happiness levels. Alternatively, we could take a non- or semi-Utilitarian perspective, arguing that in the specific context at issue, we think it is just fair to treat the two parties' satisfaction as comparable.

Either way, this assumption is enough to pin down a comparable distance for u_1 and u_2. As a next step, we can use the players' EU preferences to scale comparable preferences for intermediate divisions of

[38] Again, assuming risk-averse preferences over x. A good exercise is to work out how to adjust the statement for risk-neutral and risk-acceptant preferences.

the good. We would then have interpersonally comparable Utility functions that could be meaningfully used to ask the question, What division maximizes total (or average) Utility?

What follows if we go this route, say, by setting $u_i(0) = 0$ for both players, $u_2(1) = 1$, and $u_1(1) = a$?

Consider first the case of linear (risk neutral) preferences over the good. With $u_1(x) = ax$ and $u_2(1-x) = 1 - x$, the simple utilitarian sum is $ax + 1 - x$. This is maximized at $x = 1$ when $a > 1$, at $x = 0$ when $a < 0$, and at any $x \in [0, 1]$ for $a = 1$.

With these risk-neutral preferences, the utilitarian intuition and implication is that we maximize welfare by giving *the whole good* to the person who gets the most value out of it (and if they value it equally, any division maximizes Utility).

Exercise 14. *Solve for the naive utilitarian best in the bargaining problem with players who have risk-averse (concave) preferences, $u_1(x) = ax^{\alpha_1}$ and $u_2(1 - x) = (1 - x)^{\alpha_2}$, for the following cases.*

1. *$\alpha_1 = \alpha_2$, both in $(0, 1)$.*
2. *$\alpha_1 = 1$, $\alpha_2 \in (0, 1)$. (Get some intuition by plugging in $\alpha_2 = 1/2$, for example.)*

For $a = 1$, what is the intuition for why there is a unique solution in the case of $\alpha_i \in (0, 1)$, as compared to the risk-neutral case? What is the intuition for the why the more risk-acceptant player gets more in the socially optimal outcome (by this utilitarian welfare criterion)? How does the utilitarian optimal division in case (2) compare to the Nash Bargaining Solution in that case?

[more to come]

14 Nash equilibrium, again

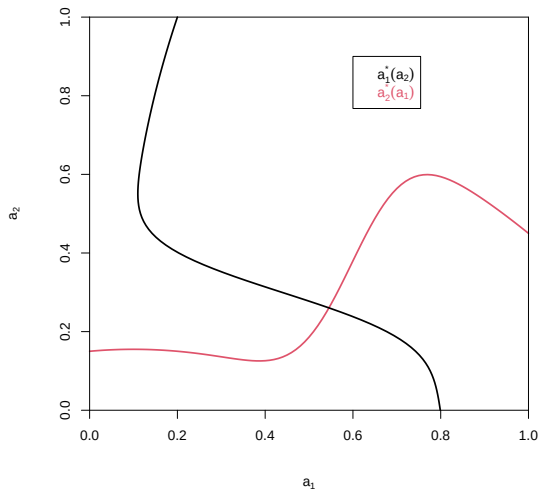
14.1 Equilibria are fixed points

For a normal-form game $\Gamma = \langle I, (A_i), (\succeq_i) \rangle$, a pure-strategy Nash equilibrium is a profile of actions $a^* = (a_1^*, a_2^*, \ldots, a_n^*)$ such that each player's action a_i^* is a best reply to the others' choices, a_{-i}^*. The primary interest in Nash equilibria is that there are good reasons to expect that any recurring "macrobehavior" is a Nash equilibrium. If not someone wants to act differently and the macrobehavior would not be stable, or recurring.

We saw that in some strategic situations there may be no pure-strategy Nash equilibrium. In a two-page note, Nash (1950*b*) observed that when the number of players and the action sets A_i are finite, there is always at least one "equilibrium point" in pure or mixed strategies when preferences can be represented by an expected utility function.

With mixed strategies, a player's best-reply correspondence becomes continuous on the set of all mixed profiles of the other players. In Matching Pennies, for instance, let α_2 be the probability that player

Figure 8: Two best-reply correspondences, illustrating that there must be a fixed point



2 plays Tails. With expected utility preferences defined over lotteries on outcomes in A, player 1 has at least one best reply for every α_2, and these best replies change "smoothly" with α_2 – there are no discrete jumps. Nash pointed out that this means that the best-reply correspondence for all players satisfies conditions of a fixed-point theorem, meaning that there has to be at least one pure or mixed profile $\alpha^* \in \times_i \Delta(A_i)$ such that $\alpha^* = BR(\alpha^*)$.[39]

Figure 8 provides a graphical intuition, which shows best-reply functions for some two-player game in which the strategy sets are $a_1, a_2 \in [0, 1]$. If it is continuous, a best-reply function for player 1 to a_2, $a_1^*(a_2)$, has to be a line that starts on the x-axis and ends touching the segment $(0, 1)$ to $(1, 1)$. Likewise, if it is continuous, a best-reply function for player 2 to a_1, $a_2^*(a_1)$, has to be a line that starts on the y-axis and ends touching the segment $(1, 0)$ to $(1, 1)$. Think about it, and you should be able to convince yourself that there is no way to draw two such continuous lines without them crossing each other at least once (as in this example). If they cross, this is a Nash equilibrium since each is choosing a best reply to the other's choice.

Exercise 15. *Draw the best-reply correspondences for Matching Pennies, on a graph akin to Figure 8. To do this, make the x-axis the probability with which player 1 chooses Heads, and the y-axis the probability with which player 1 chooses Heads. Keep in mind that in this game, each player has a mixed strategy such that the other player has more than one best reply (thus, best-reply correspondences rather than functions).*

Exercise 16. *Try to draw a continuous best-reply correspondence in a setting like Figure 8 such that the correspondence touches the 45 degree line an even number of times. What do you have to do? What do you think would happen if the underlying payoff functions were now slightly "perturbed,"*

[39] Because $BR(\alpha)$ is a correspondence rather than a function, we actually need to specify what continuity means for a correspondence. It turns out that there are two types of continuity of interest and for Nash's theorem the relevant one is called upper hemicontinuity. A BR correspondence is upper hemicontinuous at $\alpha \in \Delta(A)$ if for every sequence $\alpha_n \to \alpha$ and $\beta_n \to \beta$ where $\beta_n \in BR(\alpha_n)$, $\beta \in BR(\alpha)$. BR is upper hemicontinuous on $\Delta(A)$ if this is true for all $\alpha \in \Delta(A)$.

meaning changed by a very small amount? What does this suggest about the "typical" number of Nash equilibria?

In almost all of the models considered in what follows, we represent players as having continuous action sets – not finite. And our workhorse models of electoral politics will for convenience usually imagine a continuum of voters. So strictly speaking Nash's existence theorem won't apply. However, we almost always have adequate continuity in the payoff functions for Nash equilibria to still exist. There are plenty of theorems establishing continuity conditions on payoff functions necessary for equilibrium existence when strategy sets are continuous. When non-existence is an issue in a model, in my experience it always looks like a "technical artifact" without substantive import.

Exercise 17. *Find any and all mixed-strategy Nash equilibria of the games in Figure 1. Hint: Remember that to be "willing to mix," a player needs to get the same expected payoff from any pure strategies that she plays with positive probability in the mix. So to find player i's equilibrium mixing probability you need to ask what mixture by i makes player $j \neq i$ indifferent between j's actions.*

Exercise 18. *Fixed point theorems should not be seen as mysterious or magical. You can prove the following very simple version just by going through cases and using the intermediate value theorem for one of them. Consider a continuous function $f : [0, 1] \to [0, 1]$. Show that there must be at least one $x^* \in [0, 1]$ such that $f(x^*) = x^*$. (Hint: Consider the function $g(x) = f(x) - x$.)*

14.2 Iterated deletion of strictly dominated strategies and rationalizability

Consider a(n imaginary) Congressional district with five equal-sized blocks of voters, that we describe in terms of their political positions on a left-right dimension: Left, Center Left, Center, Center Right, Right. Suppose that two candidates, 1 and 2, simultaneously choose campaign platforms, $x_i \in \{L, CL, C, CR, R\}$, after which all the voters in a block vote for the candidate whose platform is closest to their preference. Blocks divide their votes equally between candidates that are equally far away. So, for example, if 1 chooses $x_1 = C$ and 2 chooses $x_2 = R$, the vote shares are $(3.5, 1.5)$. Likewise, if the candidates "locate" at the same position, vote shares are always $(2.5, 2.5)$.

Exercise 19. *Represent this situation as a two-player normal-form game between the two candidates, with payoffs being their vote shares. Find any and all pure-strategy Nash equilibria. Also ask whether the game has any dominated strategies. In terms of our taxonomy, what kind of strategic situation is this, from the perspective of the candidates?*

Hopefully you noticed that locating at an extreme, L or R, is never a best reply for either candidate, and thus is strictly dominated. If each candidate believes that the other is rational and has the preferences as represented, then each can anticipate that the other will not choose L or R. So in effect the problem is then reduced so that $A_i = \{CL, C, CR\}$ with the same payoffs for the remaining action combinations. And in this reduced game, C dominates CL and CR. So a rational candidate who believes the other is rational, *and who believes that the other believes that they are rational*, can conclude C must be optimal, and we are led to the prediction of (C, C).

This is an example of the solution concept known as *iterated deletion of strictly dominated strategies* (IDSDS). It is tedious but not difficult to formalize the procedure. I will just note that:

		2				
		L	CL	C	CR	R
1	L	2.5, 2.5	1, 4	1.5, 3.5	2, 3	2.5, 2.5
	CL	4, 1	2.5, 2.5	2, 3	2.5, 2.5	3, 2
	C	3.5, 1.5	3, 2	2.5, 2.5	3, 2	3.5, 1.5
	CR	3, 2	2.5, 2.5	2, 3	2.5, 2.5	4, 1
	R	2.5, 2.5	2, 3	1.5, 3.5	1, 4	2.5, 2.5

(a) In each "round" of deletion, you have to ask whether any remaining pure action is strictly dominated by any pure *or mixed* action profile, because it is possible for an action that is never a best reply to be dominated only by a mixed strategy but by no pure strategy.[40]

(b) With IDSDS, the order of deletion does not matter, you will end up with the same final set regardless (this is a theorem). Interestingly, the order can matter if you do iterated deletion of *weakly* dominated strategies, which makes IDWDS rather less appealing as a solution concept.

IDSDS has limited or zero purchase with coordination games. Think of Battle of the Sexes. Since neither player has an action that is strictly dominated, neither action is eliminated and the set of action profiles that survives IDSDS is just the full set A.

However IDSDS can have remarkably strong implications in some commitment and some strictly competitive strategic situations (note that the Downsian candidate competition model above is a strictly competitive game, in fact, constant sum).

Exercise 20. *Write down an integer in the set $A_i = \{1, 2, 3, \ldots, 100\}$. The set of people in the class who write down the smallest integer share this number of dollars equally between them. Everyone else gets nothing. Solve this game for Nash equilibrium and ask if it has any strictly dominated strategies. (Hint: It is often easier to ask if any strategy is never a best reply, in which case it is almost certainly strictly dominated.) What kind of strategic problem is this? What does IDSDS imply in this game?*

In experimental play of this game, I have found that at least one person always writes down 1, but a lot of people will write down a low number greater than 1, such as 5, or 8, or 12, and few people write down high numbers like 50. If the game is then played repeatedly, there is fairly rapid convergence towards almost everyone writing down 1, with some frustration about this and some people continuing to write down higher numbers perhaps hoping to pull others up?

In this game, $a_i^* = 1$ for all i is the unique Nash equilibrium and it is also the unique profile that survives IDSDS. In general we have the following easy theorem, which you should be able to prove just by thinking about the definition of Nash equilibrium.

Theorem 2. *If $a^* \in A$ is a pure-strategy Nash equilibrium of the normal-form game Γ, then for all i, a_i^* survives IDSDS. If $\alpha^* \in \times_i \Delta(A_i)$ is mixed-strategy Nash equilibrium of Γ, then every action that has positive probability in α^* survives IDSDS. However, it can happen that a profile $a \in A$ is not Nash but whose components survive IDSDS. The set of Nash equilibria is a subset of the profiles that survive IDSDS.*

[40] See the game in footnote 2 for an example.

At first thought, IDSDS may seem appealing as a solution concept in comparison to Nash equilibrium. This is because when we predict a specific Nash equilibrium, we are assuming that somehow the players have arrived at correct expectations about what the other players are doing, and the assumption of rationality by itself does not imply that players could arrive at correct expectations in any given strategic situation. For example, in the following Battle of the Sexes, nothing in the assumption of rationality, and the stronger assumption of common knowledge of rationality and the game, rules out that player 1 could reasonably choose D expecting that player 2 is going to choose L (which is a best reply to D), while at the same time player 2 reasonably chooses R thinking that 1 is going to choose U (which is a best reply to R). But (D, R) is not Nash. Another way to put this is that individual rationality by itself does not rule out that players might have *mistaken conjectures* about what the other players are choosing.

		2	
		L	R
1	U	$0, 0$	$20, 30$
	D	$30, 20$	$0, 0$

However, in strategic situations where IDSDS has a lot of bite, rationality and common knowledge of rationality are by themselves enough to imply that players can simply *deduce* what the other (rational) players are going to choose. This can be very compelling in strategic situations where a specific prediction a^* emerges after just one or a small number of "rounds" of deletion of dominated strategies, such as these.

		2	
		L	R
1	U	$20, 20$	$0, 30$
	D	$30, 0$	$10, 10$

		2	
		L	R
1	U	$5, 5$	$0, 3$
	D	$0, 11$	$20, 10$

By contrast, in the platform choice and "lowest number" games above, the unique IDSDS predictions emerge only after multiple rounds of deletion of dominated strategies. *Each round of deletion implies another level of common knowledge of rationality*. Consider the platform-choice game, and assume that instead of five issue positions, there are 101, at locations $X = \{0, 1, 2, \ldots, 100\}$. For me to "delete" 1 and 99, I need to believe that you are rational so you wouldn't play 0 or 100. To delete 2 and 98, I need to believe further that you think that I am rational, since otherwise I might think that you would play 99 in the expectation that I was going to choose 100. With each additional round of deletion we require another "level" of mutual knowledge of rationality ("I know that you know that I know that you know ...).

This may well strike you as implausible in some sense. It is a striking and true feature of humans that we are capable of internally representing, and we constantly handle, layers of beliefs about others' beliefs. For instance, at a dinner party you might condition your conversation on the fact that you know that A and B had a relationship; that A knows that you know this; B does not know whether you know

this; and that A does not know whether you know that B does not know this. But the formalization of "beliefs about beliefs" can get pretty artificial in the limit, as it were. The point is that in strategic situations where IDSDS has very strong implications, this may require leaning hard on the assumption of *common knowledge* of rationality, which in its formal sense can be problematically strong in some strategic situations.

IDSDS arrives at implications for rational choice in a strategic situation by asking what rational players would *not* choose, and going from there. We can also approach the question from the opposite direction, asking "what is the set of actions that a rational player could rationally choose in Γ, assuming common knowledge of rationality and Γ?" This approach leads to a solution concept known as *rationalizability*, so called because it picks out actions $a_i \in A_i$ that can be "rationalized" – that is, a rational player could justifiably choose it. The intuition: Consider a two-player game Γ. An action a_1 is rationalizable if it is a best reply to some beliefs about what player 2 is choosing, $\alpha_2 \in \Delta(A_2)$, where those beliefs put positive weight only on actions by 2 that in turn can be rationalized by a belief about player 1's strategy, where *those* beliefs can in turn be rationalized in the same way, on and on. Another way to say this is that a pure strategy a_i is rationalizable if and only if it is part of some pure or mixed Nash equilibrium profile.[41]

There are several ways to formally define rationalizable strategy sets. Here is a particularly clean one, deployed for a two-player game:

Definition 19. *$A_1^R \subset A_1$ is the set of* **rationalizable pure strategies** *for player 1 and $A_2^R \subset A_2$ for player 2 if these are the largest subsets of A_1 and A_2 such that*

1. *for all $a_1 \in A_1^R$, a_1 is a best reply to some belief $\alpha_2 \in \Delta(A_2^R)$, and*
2. *for all $a_2 \in A_2^R$, a_2 is a best reply to some belief $\alpha_1 \in \Delta(A_1^R)$.*

Since a rationalizable strategy must be a best reply to some belief about their other players' play, strategies that are never best replies are ruled out. And since a rationalizable strategy has to be a best reply to a belief that puts positive weight only on actions that can in turn be rationalized by a belief about the others' play, actions that are best replies to beliefs that rely on dominated strategies are in turn ruled out. So if you think about it it should not be too surprising that

Theorem 3. *The set of rationalizable actions for i in Γ is identical to the set of actions that survive IDSDS. It is also identical to the set of actions that are played by i with positive probability in some Nash equilibrium of Γ.*

[41] For the full formal development, see Avi's notes or any good game theory text. The solution concept of rationalizability was developed by Pearce (1984) and Bernheim (1984).

Part III

Good policy, bad policy

15 Interpersonal comparisons of utility (are ok)

15.1 Interpersonal comparisons

We have seen that ordinal utility functions encode information only about how a person ranks a set of outcomes in terms of preference. A cardinal utility function encodes additional information about the intensity of a person's preferences over the outcomes, where "intensity" is understood in terms of – indeed, is defined by – preferences over gambles or lotteries on the outcomes.

Neither ordinal nor cardinal utility functions encode any information about how much one person likes an outcome $a \in A$ as compared to another person. If I tell you that Bob's utility for a cup of tea is 3 while Marilyn's utility for a cup of coffee is 7, this tells you precisely *nothing* about who gets more enjoyment from their preferred beverage (unless I have made some other assumptions).

We might have hoped that cardinal utility would have helped us out here, since it is in one sense measuring "intensity" of preference. But it does not. Recall that if the cardinal utility function u represents a person's preferences over the lotteries $\alpha \in \Delta(A)$, then so does $v = au + b$ where $a > 0$ and b is any number. We can arbitrarily scale the numbers up or down without affecting their ordering and relative spacing, which means that we do not have any common reference points that would allow us to compare "levels of happiness" or wellbeing across people.

Consider the following four statements.

1. Bob prefers tea to coffee, whereas Marilyn prefers coffee to tea.
2. Bob has a more intense preference for tea, relative to coffee and water, than Marilyn has for coffee, relative to tea and water.
3. Marilyn's gain in satisfaction from coffee relative to tea is smaller than Bob's gain in satisfaction from tea relative to coffee.
4. Marilyn gets a greater sense of wellbeing from a cup of coffee than Bob gets from a cup of tea.

All four are interpersonal comparisons of a sort, but only the third and fourth are what the literature means by an "interpersonal comparison of utility." The first is just expressing a difference in ordinal rankings, and the second tells us something about gambles that Bob and Marilyn would be willing to run on a set of common outcomes. The third and fourth introduce a new thing, which is an implicit idea of wellbeing or happiness that can be compared across people. As we will see, the fourth claim assumes more about what can be compared than the third.

Economists often give the impression that there is something wrong or illegitimate about even asking whether Bob gets more happiness from a cup of tea than Marilyn gets from a cup of coffee. You hear it said that "you can't make interpersonal comparisons of utility." But of course you can! This is wildly wrong. We do it literally all the time. Further, politics as a human activity is incomprehensible if we do not see that it is to a great extent *about* making interpersonal comparisons of utility.

To give a simple example, suppose Bob and Marilyn are a couple, and they happen to find themselves in a situation where they have only enough change for one coffee or one tea. Or perhaps they are choosing a restaurant to go to. It is completely normal for them to be able to assess which is likely to be the "utility maximizing" choice for the collective that is Bob and Marilyn, even if their first preferences are different. They may know, for example, that Bob strongly dislikes coffee while Marilyn slightly prefers coffee but doesn't really care that much. This is an interpersonal comparison of utility.

In terms of a numerical representation, it is as if

$$u_B(\text{tea}) - u_B(\text{coffee}) > u_M(\text{coffee}) - u_M(\text{tea})$$

is a meaningful comparison, and therefore also

$$u_B(\text{tea}) + u_M(\text{tea}) > u_B(\text{coffee}) + u_M(\text{coffee}). \tag{2}$$

The choice of tea makes for greater "total utility" (or equivalently here, average utility) for the Bob and Marilyn collective. By what I will call the *naive utilitarian criterion*, tea is the best policy choice in this example.

What additional information needs to be encoded in a cardinal utility function for this interpersonal comparison to be meaningful? We saw that with cardinal utility, if u_i represents a person's preferences then so does $v_i = a_i u_i + b_i$ for any $a_i, b_i \in \mathbb{R}$ with $a_i > 0$. So consider (2) for the equivalence class of cardinal representations:

$$a_B u_B(\text{tea}) + b_B + a_M u_M(\text{tea}) + b_M > a_B u_B(\text{coffee}) + b_B + a_M u_M(\text{coffee}) + b_M.$$

Rearranging,

$$a_B(u_B(\text{tea}) - u_B(\text{coffee})) > a_M(u_M(\text{coffee}) - u_M(\text{tea}))$$

Notice that b_B and b_M drop out of the comparison. You can think of b_B and b_M as baseline levels of wellbeing that are independent of, or unaffected by, the choice at issue. Perhaps Bob is in a better mood than Marilyn or perhaps Marilyn is just generally happier than Bob. The part of their sense of wellbeing or happiness that does not depend on the choice of coffee or tea is irrelevant to the question of whether tea or coffee maximizes average wellbeing.

We would need information about baseline-level wellbeing to make statements like (4) above, but not to make an interpersonal comparison like (3). This is fortunate. As the inequalities above show, (3)-type claims are what we need to answer the question of which outcome or policy maximizes average sense of wellbeing. In other words, we don't need to be able to measure individuals' "utility temperature." We need to be able to compare across people the *change* in satisfaction from one outcome to another,

which is like comparing *differences* in "utility temperature" across people (not levels).[42]

And with preferences representable by a cardinal utility function, we only need to fix one such comparison for each pair of individuals. Consider the policies $A = \{\text{coffee, tea, water}\}$, assuming as before that Bob has tea $\succ_B$ coffee $\succ_B$ water and Marilyn has coffee $\succ_M$ tea $\succ_M$ water. Because the baseline level of wellbeing is not relevant for our question, we can conveniently set both of their utility numbers for water to be zero: $u_B(\text{water}) = u_M(\text{water}) = 0$.[43] If we, or Bob and Marilyn, can say something like "Bob's gain in sense of wellbeing for coffee versus water is x times as large as Marilyn's gain for tea versus water," then we can

- first, set $u_M(tea) = 1$ and $u_B(\text{coffee}) = x$, using the information about the interpersonal comparison, and then
- second, set $u_M(\text{coffee}) = \alpha > 1$ and $u_B(tea) = \beta > x$ according to their willingness to accept gambles on A.

This hypothetical procedure leads to a set of interpersonally comparable utility functions as in Table 15.1.[44] They represent just enough information about preferences that the sum of utility numbers across outcomes implies a social ordering from best to worst policy.

The ethical standard here is avowedly utilitarian. The best policy is said to be the one that maximizes "wellbeing" where wellbeing is understood in terms of the agents' own preferences and degrees of "satisfaction" or "happiness." Individuals' welfare is weighted equally in the overall reckoning.

In the sections that follow we will constantly use this kind of naive utilitarian approach when assessing whether equilibrium outcomes in political economy models are more or less messed up. "Naive," because utilitarianism as an ethical theory should not be taken too seriously. It treats preferences and causes of happiness like tastes that are not to be disputed. But no sane person thinks that any and all preferences should be respected. Section 16 takes up the issue of moral disagreements, which we need to consider in order to assess what is good or bad policy in common situations where some people strongly disagree that other peoples' policy preferences merit respect and weight.[45]

I have no doubt but that you are wondering how, as a practical matter, it would be possible discover the relationship between two peoples' sense of welfare gain or loss between two outcomes. That is, what is x in $u_i(a) - u_i(b) = x(u_j(a) - u_j(b))$ for two outcomes $a, b \in A$ and two people i and j?

Marilyn and Bob can make such judgements because they know each other well and have enough care for each other's happiness that they will accurately report how strongly they feel about one option

[42] Sen (1970) ...

[43] Make sure you understand why this does *not* mean that Bob and Marilyn get the same "hedonic utility," or sense of wellbeing, from water.

[44] We could just as well have asked about the relationship between Bob's gain from tea versus water and Marilyn's gain from coffee versus water, and then drawn on preferences over gambles to pin down the utility number for the third outcome.

[45] In addition to its absurd refusal to pass judgement on individual preferences – which amounts to a complete abdication for a moral theory – naive utilitarianism also faces a lot of reasonable criticism on the grounds that summing utilities implies indifference to the distribution of welfare across people, and thus is indifferent to ideas of fairness. Sen (2017, chapter A3 and elsewhere) is an excellent treatment.

Table 2: An example of utility functions that can be meaningfully summed across outcomes

	$u_{Marilyn}$	u_{Bob}
coffee	α	x
tea	1	β
water	0	0

Notes: $\beta > x > 0$, $\alpha > 1 > 0$. See text for the construction.

or another (restaurants, beverages, etc.). Of course neither condition holds so strongly in matters of politics, which makes identifying approximately "welfare maximizing" policies more difficult.

Nonetheless, the push and pull of politics is substantially made up of people expressing levels of happiness or unhappiness through their actions ("political behavior"). These expressions, together with the capacity for empathy for others' wellbeing, form the basis for interpersonal comparisons of welfare. Political institutions are often naturally understood as mechanisms for eliciting preferences so as to do better on welfare maximization or avoidance of bad outcomes like war.

As discussed further in Section 17, for some policy questions it can make sense simply to *stipulate* a common value across people, even when we know that there is preference variation across people that is hard to elicit truthfully. For example, in that section we consider the consequences of stipulating that the gain in wellbeing associated with going from dire poverty to having basic needs met is similar across people and households.

15.2 Arrow's theorem is not relevant as it is often interpreted

Suppose we know how each of n people in a group or society rank orders policy options in a finite set A that has at least three options. Arrow (1963) asked if it was possible to use these individual preference orders to construct a *social welfare functional* (SWF) that orders the policy options from best to worst while respecting the following conditions:

- (Pareto) If everyone has $a \succeq_i b$ in their individual rankings, so does the social ordering.
- (Universal domain) The SWF produces an ordering for any and all sets of individual orderings.
- (Independence of irrelevant alternatives, or IIA) The ranking of a and b in the social ordering depends only on how individuals rank a and b, and not on the position of other alternatives with respect to a and b or other alternatives.
- (Non-dictatorship) There is no individual i such that $a \succeq b$ in the social ordering if and only $a \succeq_i b$, for any $a, b \in A$.

Arrow proved that no SWF that can satisfy all four conditions, a surprising and even shocking result given that the conditions seem at first glance both desirable and unexceptional.

Arrow's result is often interpreted as implying that the idea of a best policy for a collective is incoherent or meaningless. Instead, it shows that if we only admit information about how individuals rank policies, we don't have enough to produce a social ordering that has desirable features.

IIA is particularly problematic regarding what it disallows. Suppose that a bare majority just barely prefers policy Tea to Coffee, and the bare minority just slightly prefers Coffee to Tea. With these preferences, a reasonable SWF might rank Tea ahead of Coffee. Now imagine that we change the preferences of the minority who prefer Coffee to Tea so that they all *strongly* prefer Coffee to Tea. IIA requires that the SWF continue to rank Tea ahead of Coffee for the society, since no one has changed their individual ranking of these two options. IIA thus does not allow any comparison of intensity of preference, or change in well-being from one policy to another, across individuals to influence the collective ordering.

Both in the 1950s and still today, Arrow and other economists were strongly under the influence of the logical positivist school of thought about Science. The logical positivists were *extreme* skeptics. They maintained that statements about entities that are not directly observable are meaningless. For example, in B.F. Skinner's application of the approach to psychology, it is meaningless to talk about minds, since we can only observe "behavior." Beginning in the 1920s and 1930s (the heyday of logical positivism), economists moved away from the utilitarian tradition of talking about utility as reflecting an internal state. Instead, they said that choices are what is observable and "utility" is just an analytic convenience that summarizes information about "revealed preferences" (derived from observing choices). This can be a reasonable and useful move in terms of making it clear how little one needs to assume for certain implications to follow in particular strategic settings.

But in retrospect it seems odd that internal states such as one's subjective sense of wellbeing were viewed as not directly observable. The great skeptic Descartes had argued that these are basically the *only* things that are directly observable. And while philosophers can debate how do I know you all aren't clever machines that don't have subjective experiences anything like mine, this hardly seems relevant or plausible enough to deny us the possibility of meaningful interpersonal comparisons when choosing social or economic policies.[46]

15.3 Gibbard and Satterthwaite's version is

While I think it is a mistake to interpret Arrow's theorem as ruling out meaningful statements about better or worse policies for a collective, Gibbard (1973) and Satterthwaite (1975) showed that its logic has a remarkable implication for any system that tries to *identify* good policies based on individuals' reports of their preferences.

Imagine an automated Government Machine, which takes as inputs individuals' reports of their prefer-

[46]Besides philosophy of science, a related reason for the field's sidelining of interpersonal comparisons was the desire to make Economics an objective, scientific field that could advise on matters of policy without taking positions on values or moral philosophy. Robbins (1938) is sympathetic to utilitarianism and the idea of "equal satisfaction" across people, but wants to separate positive arguments about, for instance, the effect of a tax on prices and quantities, from normative arguments of political philosophy.

There is a long history of responses to Arrow's theorem that reject its rejection of interpersonal comparisons of utility. [cites, Sen, Harsanyi, Gibbard, . . .]

ence orderings over policies in a set A (with $|A| > 2$). In one important application, the "policies" are candidates for office and the Government Machine is a voting rule (for instance, first-past-the-post, or ranked choice).

For any set of reports, the Machine spits out a policy choice. Gibbard and Satterthwaite showed that if the Government Machine's decision rule respects Arrow's conditions above, it must be the case that there are situations in which individuals would optimally misrepresent their true preferences. With more than two alternatives, there is no "non-manipulable" (Gibbard) or "strategy-proof" (Satterthwaite) aggregation rule that respects Pareto, IIA, and non-dictatorship for any set of individual preferences. Admitting cardinal utility does not materially change this result.[47]

We often believe that policy should depend in some way on people's preferences, in situations where it is not clear what peoples' preferences are. The Gibbard-Satterthwaite theorem points us in the direction of a very general – and I think substantively important – explanation for why things are messed up.

Namely, optimal policy depends on peoples' preferences, but because people have private information about their preferences and because they anticipate that they may have disagreements, there is not in general any great cost-free way of discovering true preferences in order to make social decisions. There is no good way to design a rule (or "institution") such that people would always want to truthfully reveal their preferences about what should happen. To get truthful revelation – which we want in order to be able to choose a good policy – there are going to have be costs and benefits applied to particular statements or actions. This is the province of the field known as *mechanism design*, which we will encounter in various examples in what follows.

To make the point more concretely, consider the majority-rule mechanism. If a majority has a slight preference for Tea over Coffee, but the minority has a strong preference for Coffee over Tea, we might think (implicitly using an interpersonal comparison of utility) that the best policy is Coffee. But how can we learn how much the minority cares? If saying "I am passionate about Coffee versus Tea" was enough to get the result, then even a person with a weak preference would have an incentive to so report. In real-world politics, we see minorities expending costly effort of all sorts, up to and including violence, to try to credibly convey the intensity of their preferences.

16 Good policy: Maximize welfare or do the right thing?

16.1 Moderation in all things

Suppose that the "collective" of Bob and Marilyn faces a decision about what policy to choose. The options are Coffee, Tea, and Water, which as above are meant to stand in for all manner of real-world issues. For instance, it could literally be whether the National Drink should be coffee, tea, or water, or it could be whether eating meat and fish should be legal, meat illegal but fish legal, or neither meat nor

[47] Hylland (1980) showed that with cardinal utility, the only non-manipulable rules that satisfy the Arrow conditions are "random dictatorships" – chose a random individual and have this person make the social decision. See also Dutta, Peters and Sen (2007). To be honest, random dictatorship – or better, selecting a small group at random and having them deliberate, as with juries – has some attractive features. On such methods of *sortition*, see Stone (2011) and Elster (1989).

fish legal.

Suppose that Bob and Marilyn both prefer the policy Coffee to Tea and Water. From the ethical perspective that says it is a good thing to maximize wellbeing as people perceive this for themselves, this is an easy case for the question of what is the best policy. Coffee, the Pareto-optimal outcome.

But what if there are strong arguments that the Coffee policy is highly immoral and wrong? You can surely give examples of hypothetical policies that you think would definitely *not* be good policies even if every person in the country or district in question happened to believe were great. There are many examples of historical policies that people used to think were fine and good, but that we now think were terrible and wrong.

Suppose next that Bob prefers Tea to Coffee and Marilyn prefers Coffee to Tea. Both prefer either to Water. What is the best policy for the collective?

If the actual policies concern something like What should be the National Drink?, then Bob and Marilyn might well agree that it is appropriate that the best policy is the one that maximizes total happiness, or wellbeing. As suggested by the analysis of the last section, if Marilyn is really passionate that Coffee should be the national drink but Bob doesn't care that much, then the best policy would be Coffee. Here, a *welfarist* criterion – and even a specifically *utilitarian* criterion – for choosing the best policy seems appropriate and defensible.

It is less clear what the right policy should be in cases where Bob, Marilyn, or both have the view that the other's most preferred policy is simply wrong, immoral. It is not clear why "choose the policy that maximizes subjective perception of wellbeing" should have force for the loser, or indeed for anyone if we allow that what is right may not be what maximizes happiness (understood in terms of preference). Why should I say "Well, I think the policy you prefer is fundamentally immoral but you really, really like it, so ok that's what should happen"?

In such situations, which are completely standard in problems of policy choice, Bob and Marilyn

- can argue and deliberate,
- they can strike a provisional compromise, or
- they can try to use force to coerce each other.

On both welfarist and do-the-right-thing grounds, violence and civil war should and will often be seen as terrible outcomes by pretty much everyone.

Let Water be civil war, an outcome that no one likes at all but that results, or may result, if Bob and Marilyn cannot agree on some deal concerning Coffee and Tea. Even if they cannot manage to argue and deliberate to a new moral consensus, there is a strong argument that some kind of provisional compromise would be a good policy outcome in such situations. For example, alternate between Coffee and Tea, or choose a compromise policy that is intermediate on more continuous dimensions of whatever is at issue (for instance, how restrictive or not restrictive should policies on gun control or abortion be).

The example suggests a non-utilitarian argument in favor of moderate policies as (provisionally) normatively best. People often disagree about what is the right policy, and often on moral grounds in the specific sense that they would not accept as relevant the criterion of maximizing "total utility" under current preferences. Implementing a policy that a large majority dislikes or that a minority intensely dislikes because it is viewed as immoral risks social peace, a commonly valued and important component of wellbeing. So there is a case for provisional compromise, moderation, to allow persuasion and argument as people work through the moral issues over time.

The argument is non-utilitarian because it allows that what makes a policy best may have to do with its moral qualities rather than whether it maximizes wellbeing given current preferences. Current preferences get respect here as a pragmatic matter, for the sake of the common value of social peace.

16.2 A social peace function

So far as I know, arguing for moderate policies on the grounds that they conduce to social peace when there are moral disagreements over what is right is not a typical move in either recent normative economics or moral philosophy. The former is focused on justifications for different social welfare functions, and a standard premise is that everyones' preferences merit some respect whatever they are. In moral philosophy and ethics, the concern is mainly with what is right or wrong in specific issue areas or policy questions.[48]

Without intending anything like an extended development, which I don't have, here is a suggestion about how one might proceed.

A group of n people consider what policy or policies to choose from the set A. Suppose there is one outcome, a_0, that everyone agrees is the worst in A. Call this "civil war." And for each person i, let a_i^* be her most preferred outcome in A (or one from the set of her most preferred outcomes).

If the members of the group have preferences on A representable by expected utility, we can next scale their utility functions so that $u_i(a_0) = 0$ and $u_i(a_i^*) = 1$. Now $p_i(a) = u_i(a)$ has the interpretation that this is the probability at which i is indifferent between the policy a and a gamble that gives her favorite social outcome with probability $p_i(a)$ and civil war with probability $1 - p_i(a)$.

Thus if $p_i(a)$ is close to 1, this means that i is quite happy with a relative to her most preferred policy and civil war, and so would be relatively *content* with a as the policy, in terms of motivation to change things at risk of conflict. If $p_i(a)$ is close to zero, this would indicate a policy that i thinks is very bad (relative to her most preferred and civil war) and would be willing to run a large risk of violent conflict to not have to live with.

The average $p_i(a)$ is then a measure of the degree of social contentment with a, understood in terms of average motivation or willingness to risk costly conflict to change it. We could define a "social peace function"

$$SPF(a) = \frac{1}{n} \sum_i p_i(a),$$

[48]For great examples of work in normative economics, see Roemer (1996), Adler (2019), and, farther back, Sen (1970).

and choose a policy a that maximizes this as a proposal for what would be socially best.

This may look utilitarian in that we are adding up utilities, but it is not. We never made an assumption about interpersonal comparability and we don't need to. What is being compared is willingness or motivation to risk a mutually-agreed very bad outcome against changing the policy for something each person views as better.

When we scale the utility functions to $[0, 1]$, we are in effect discarding (or not needing) information about interpersonal differences in wellbeing between any two policies a and a'. As a result the policy that maximizes $SPF(a)$ using $u_i(a)$ need not be the policy that would maximize a utilitarian social welfare function with interpersonally comparable preferences $v_i(a)$, where $u_i(a)$ is the affine transformation of v_i to $[0, 1]$.

Table 16.2 illustrates. Coffee is the best utilitarian policy here, since $1.5 + .6 > 1 + 1$. But Tea would be best for social peace since $1 + .6 < .67 + 1$. Even though Marilyn gains more wellbeing from Coffee versus Tea than Bob does from Tea versus Coffee, relative to the bad outcome of Water she "cares less" than Bob. If Bob and Marilyn have a principled, moral disagreement about Coffee versus Tea, then it is not clear why the best policy should be the one that maximizes a subjective happiness standard. There may, however, be a *pragmatic* justification for policy that reduces risk of a mutually agreed bad, while attempts at persuasion and perhaps collection of more evidence continue.[49]

Morally speaking, perhaps neither Bob nor Marilyn should use the threat of Water to try to get their way on Coffee versus Tea. But this is inevitable in real politics, where situations that have the strategic structure of bargaining problems are truly ubiquitous.[50] And some policy needs to be chosen in the face of moral disagreements. They need to make some choice.

Exercise 21. *Suppose that the group of people have preferences over a policy dimension that we represent with the interval $X = [0, 1]$. To be discussed at much greater length below, you can think of the dimension as a Left-Right, liberal-to-conservative scale for a policy on some specific matter. Individuals have a most preferred policy $x_i \in X$, and preferences represented by the utility function $u_i(x) = -(x_i - x)^2$. There is also a disagreement outcome that occurs if they can't agree on a policy in X, which could be civil war. All have utility $-d$ for this outcome, with $d > 1$.*

Find $p_i(x)$, the probability that makes i indifferent between policy $x \in X$ and the gamble between her most preferred policy and civil war. Then identify the policy that maximizes the social peace function as defined above.

As this example illustrates, the naive utilitarian and social peace standards (as defined above) coincide when we take the welfare gain from going from the worst outcome to each individual's most preferred outcome to be the same. This assumption is often made in work on the questions considered in the next section, concerning how best to distribute income or material goods using tax policy.[51]

[49] Normally, policy options are not simple and binary but concern sets of choices that can be ranged on a continuum over which the parties have conflicting preferences. For instance, there could be a set of policies a_k, $k = 1, 2, \ldots, m$, with a_1 = Coffee, a_m = Tea and $k < l \Rightarrow a_k \succ_{Marilyn} a_l$ but $a_l \succ_{Bob} a_k$. Now a moderate policy will be best from this social peace perspective whenever there is a k greater than 1 and less than m such that $u_M(a_k) + u_B(a_k) > 1$.

[50] Recall the definition of a bargaining problem in section 9.6; the Coffee, Tea, or Water problem posed here is bargaining

Table 3: Coffee, tea, and social peace

policy	Marilyn		Bob
a	$v_M(a)$	$u_M(a)$	$v_B = u_B(a)$
Coffee	1.5	1	.6
Tea	1	2/3	1
Water	0	0	0

Notes: $v_i(a)$ is an interpersonally comparable representation. $u_i(a)$ is v_i scaled to $[0, 1]$.

17 Distributive justice: Taxation for redistribution and public goods

17.1 . . . jedem nach seinen Bedürfnissen

Suppose that we have a lot of money, say $1 million, and our task is to distribute it among a set of people or households, named $i = 1, 2, \ldots, n$. Our goal is to allocate the money in the most ethically defensible way possible. This is a stylized problem of "distributive justice."

If we know nothing about the individuals – and if we can assume that they all like more money rather than less and do not care how much others get – then arguably we should give everyone the same amount, $1/n$. Without any information to distinguish between them, there is no basis for deviating from equal shares. And since everyone likes more rather than less, it is probably not good to "leave wellbeing on the table" by not allocating the full $1 million (the Pareto criterion at work).[52]

But what if we know something about different individuals, such as their level of wealth? Now a

problem in that sense.

[51] More notes on $SPF(a)$: The idea of deriving a social welfare function after assuming there is a mutually-agreed worst outcome a_0 and introducing some axioms (as in Arrow's approach) appears in the social-choice literature in relation to the idea of the "Nash social welfare function." See Roberts (1980) and Kaneko and Nakamura (1979), who refer to antecedents back to Luce and Raiffa (1957). This is $NSWF(a) = \Pi_i u_i(a)$, the product of the group members' expected utility functions defined over the set of outcomes $a \in A$. The maximum of $NSWF(a)$ is the same for any positive affine rescaling of the u_i's, so like $SPF(a)$ it has the feature that it does not use or require information about interpersonal comparisons. The maximum of $NSWF$ is the n-person generalization of the Nash Bargaining Solution (NBS) for two people, to be discussed in Section X. As a positive prediction about bargaining outcomes, the NBS can be motivated as the outcome that equally balances the bargainers' willingness to risk "no deal" by demanding more. This is parallel to motivating $SPF(a)$ in terms of minimizing average motivation to risk of a_0, but the product form effectively puts far more weight on not making any one person very disgruntled. For instance, suppose one person in the group likes Coffee best and has $u_i(\text{Tea}) = \epsilon \approx 0$, while the other $n - 1$ people like Tea best and have $u_j(\text{Coffee}) = p \approx 1$. $NWSF$ selects Coffee as the best policy whenever $\epsilon < p^{n-1}$, so even for quite large n Coffee can be $NSWF$-best, oddly. In two-person bargaining it makes sense that either person can say No, upsetting the apple cart, but not so much with a large society.

[52] Notice that to conclude that it is Pareto efficient to distribute everything, we need an assumption that rules out people caring too much about what other people get – they can't be too envious. Imagine a set of people who are all desperately spiteful, so much that they all prefer getting nothing to anyone else getting anything. Then not distributing anything could be Pareto efficient even if they all prefer getting more to less, holding others' allocations fixed. The ethical stance that a good person should not be envious is an implicit assumption in a great deal of microeconomic theory, where a standard assumption necessary for the "welfare theorems" is that individuals care only about they get and not what anyone else gets.

common moral intuition holds that it is ethically better to give a larger share the lower a person's wealth, perhaps aiming to come as close as possible to equalizing final levels of wealth.

This moral intuition might be justified in several ways, both utilitarian and non-utilitarian. For example, the psychologist Abraham Maslow (1943) proposed a "hierarchy of needs" with food, water, shelter, and safety being most fundamental. The implication, consistent with almost everyone's sense of things, is that, on average, increasing the wealth or income of a household in poverty causes a greater increase in its wellbeing than would increasing the wealth or income of a very rich household by the same amount. Both Maslow's argument and this common moral sense take a position on an average interpersonal comparison of utility.[53]

We can formally represent the idea that increments to wealth matter less at higher levels of wealth with a risk-averse cardinal utility function $u_i(x_i)$ defined on amounts of money $x_i \geq 0$. From either the utilitarian or the social-peace perspective, there is no loss in generality in setting $u_i(0) = 0$ for all individuals. Suppose that we can tax and redistribute whatever wealth anyone has, with no losses or other problems, and that total wealth is $Y > 0$. Then the naive utilitarian problem is to choose a distribution $(x_1, x_2, \ldots, x_n)$ that maximizes $\sum_i u_i(x_i)$ such that $\sum_i x_i \leq Y$.

As discussed above, for this maximization problem to make any sense we need to assume that changes in the u_i scales are interpersonally comparable. For any two people i and j and any two amounts z and z', we can meaningfully compare $u_i(z') - u_i(z)$ to $u_j(z') - u_j(z)$ as reflecting the change in wellbeing that each person experiences in going from z to z'.

To implement this naive utilitarian solution, then, we would ideally need information about the subjective sense of welfare gains each individual experiences over all changes in wealth or income, an extremely tall order even putting the problem of incentives to misrepresent to the side (section 15.3).

Marx – who said "to each according to his needs" – and Maslow might argue in reply that either:

1. Ethically, what anyone makes of their allotment x_i in terms of happiness is their problem. But objectively and on average, gains in wellbeing are greater at lower levels than higher levels. So for the sake of choosing the best policy, it is appropriate to fix $u_i(z) = u(z)$ for all i, and to assume u is concave. Or,

2. As a practical matter, people are basically similar so that the individual differences in gain in sense of wellbeing from going from, say, \$20k per year to \$40k per year are not large enough to make a fuss about. And we have no good way of measuring this in any event. So let's ask what is best if we stipulate $u_i = u$ for all i. Further we do have excellent reason to think that the average gain in wellbeing for going from, say, \$10k to \$50k is much greater than going from \$1.01 million to \$1.05 million. So assume u is concave.[54]

Either way the problem then becomes to choose a feasible distribution to maximize $\sum_i u(x_i)$.

[53] Dwyer and Dunn (2022) find evidence of diminishing reported happiness returns based on random assignment of cash in an experiment (subject to assumptions about the relationship between subjective happiness and reported happiness).

[54] Some evidence for this is that people with \$1 million may not even notice when their stock portfolio changes by \$10k, but an additional \$10k for a poor person can be life-changing.

Here is an easy proof that the solution is to give everyone the same amount, Y/n. First of all, it can't be a maximum to redistribute less than the total Y, since we always increase the sum by distributing more. Second, it can't be a maximum if any one person is getting more than another – say $z' > z$ – since then we could transfer a very small amount from the one getting z' to the one getting z. Since u is concave, $u'(z) > u'(z')$, which means that the poorer person's increase in wellbeing is greater than the richer person's loss, and the utilitarian sum therefore increases.

So far, then, both a utilitarian and a social-peace perspective suggest equal distribution of material goods as a good distributive policy, or principle of distributive justice.

[*Math digression*. This is a nice example for illustrating the use of the Lagrangian for maximization with a constraint. The Lagrangian is

$$\mathcal{L} = \sum_i u(x_i) + \lambda(Y - \sum_i x_i),$$

which leads to n derivatives $\partial\mathcal{L}/\partial x_i = u'(x_i) - \lambda$. Thus we have n first-order conditions with $u'(x_i) = \lambda$, implying that at the maximum (second-order conditions are all negative) $x_i^* = x^*$, the same for all x_i. The derivative of $\mathcal{L}$ with respect to the Lagrange multiplier λ gives us back the constraint with equality as its first-order condition. This certainly binds, since we can always increase $\mathcal{L}$ by distributing more. So $\sum x^* = Y$ implying $x^* = Y/n$.

Exercise 22. *Analyze the maximization problem allowing people to have different preferences, $a_i u(x_i)$ where $a_i > 0$ and u is concave with $u(0) = 0$. Interpret in ordinary language the meaning of differences in a_i and how these affect how much a person should get in the naive utilitarian distribution.*

]

17.2 Jeder nach seinen Fähigkeiten ...

If Y is the group's total income (or output), it has to come from somewhere and usually this is going to involve effort devoted to production – work. If work is hard, or trades off against something nice like leisure, then redistribution that equalizes incomes after production may undermine incentives to produce. What to do?

In small-scale societies with simple modes of production such as hunting and gathering, group members closely observe each other's efforts and also understand each other's abilities and skills very well. Hunter-gatherer societies have also been shown by research in anthropology to be *extremely* egalitarian and to almost universally practice near-perfect sharing of total production. Neither individual hunters and gatherers nor their closest kin have any special claim to their kill or yield. Hunter-gatherer groups seem to approximate Marx's communist ideal of "from each according to his ability, to each according to his need."[55]

With more complex production technologies, monitoring effort and estimating ability becomes much harder.

[55] Boehm (1999) describes extant hunter-gatherer societies as "uniformly egalitarian."

In this section we consider a model in which individuals (or households) choose how much effort to exert, or labor to supply. Individuals differ in their ability or skills (Fähigkeiten), so that some produce more with a given amount of labor than others. Effort is directly costly, perhaps in part because there is a tradeoff with "leisure." The policy question will then be how to redistribute output after it has been produced, so as to maximize average wellbeing using concave preferences and our naive utilitarian standard.

We will start by asking what the best outcome would be if the players or a "social planner" could monitor and commit to individual effort levels. Then in section 17.4 below we introduce information and monitoring constraints and ask what is the best redistributive policy given these.

Suppose that $n > 1$ individuals (or households) produce output with a linear production function, $y_i = a_i e_i$, where y_i is output, $a_i > 0$ is "ability," and $e_i \in [0, 1]$ is effort. To start off, suppose that disutility for effort is the same for everyone in our model economy, and increases quadratically in effort, as $be_i^2/2$, where $b > 0$ scales the unpleasantness of effort relative to consumption. We constrain the range of effort to $[0, 1]$ to represent the idea that there is only so much work a person can do in a day, or whatever unit of time we are considering.

A policy choice is a list of taxes or transfers $t = (t_1, t_2, \ldots, t_i, \ldots, t_n)$, where t_i is the transfer to or from person i after they have produced output $a_i e_i$. Positive t_i is a tax and negative t_i is a transfer. We require a balanced budget, so $\sum_i t_i = 0$, and that no one can give more than their total output ($t_i \leq a_i e_i$ for all i).

Putting things together, i's payoff is $u(a_i e_i - t_i) - be_i^2/2$ and the naive utilitarian social optimum is at the choice of effort levels e_i and taxes/transfers t_i that maximize

$$\sum_i u(a_i e_i - t_i) - be_i^2/2.$$

Keep in mind that we are not yet asking how the people in this model would separately and independently choose effort levels e_i in a normal-form game. Instead we are starting off by asking, If we or the society could somehow arrange to get everyone making the socially optimal effort level, what would this be, given that people can differ in their abilities or skills?

For any given effort levels $e = (e_1, e_2, \ldots, e_n)$, we know from the last section's analysis that with concave preferences u, to maximize this sum we will want to choose transfers t so as to equalize everyone's income, so that $a_i e_i - t_i$ is the same for all. Given e, total production is $Y = \sum_i a_i e_i$ and after redistribution everyone then has Y/n to consume. So our problem becomes to choose e to maximize

$$\sum_i u\left(Y/n\right) - be_i^2/2 = nu\left(Y/n\right) - (b/2)\sum_i e_i^2.$$

Taking the derivative with respect to e_i, we find that the sum is increasing in e_i when $a_i u'(Y/n) \geq be_i$. Concavity of u means that $a_i u'(Y/n)$ is decreasing in e_i, so at a maximum we have $e_i^* = \min\{a_i u'(Y/n)/b, 1\}$. Since $u'(Y/n)$ is the same for everyone, it follows that there is a threshold ability level $\hat{a}$ such that $e_i^* = a_i u'(Y/n)/b$ for all $a_i < \hat{a}$, and $e_i^* = 1$ for all $a_i \geq \hat{a}$.

In words, ideally we would have higher ability or skilled people exert more effort, and the most skilled

work as much as possible. The intuition: Suppose two people are exerting the same effort level but one is much more productive than the other. We can shift a small amount of effort from the low-ability to the high-ability type, which has no net effect on total disutility from effort but does increase total output to be shared. Note that this logic does not apply if the high-ability type is already optimally working the full amount, $e_i^* = 1$.

At low levels of income, as with hunter-gatherers, perhaps all but those who are sick or disabled ($a_i \approx 0$) would have $e_i^* = 1$, because value of increased income relative to effort or leisure would be large (high $u'(Y/n)$).[56] At higher levels, if we have two people who optimally get to "consume leisure" ($e_i^* < 1$), the more productive one is asked to work more. In this situation, the utilitarian social optimum imposes a burden on those with higher ability. They are obligated to work more because their labor produces greater social benefit ("from each according to his ability").

This does not necessarily mean that those with higher ability would be less happy. Recall that we have normalized "baseline" wellbeing so that $u(0) = 0$. It could be that higher ability associates with greater status, in which case they would not necessarily feel exploited by the fact that they get equal consumption but are asked to work more.[57]

Exercise 23. *As a point of comparison, suppose that each person chooses her effort level separately and consumes only what she produces – there is no redistribution, or sharing. Then i wants to maximize $u(a_i e_i) - b e_i^2/2$. Consider the case of preferences represented by $u(z) = \sqrt{z}$, and allow e_i to be any number greater than zero so as to avoid corner solutions at $e_i = 1$.*

Solve for optimal e_i^* in this case and in the social optimum case just considered. Calculate equilibrium total output in the two cases and compare.[58]

You should find that as long as there are differences in ability across people, total output is higher under the utilitarian optimum than if people worked and consumed separately, with no sharing. And this is true even though we have been using a completely individualistic production technology, in which a person's output depends only on their own effort and not on anyone else's.[59]

Of course, in reality complementarities in production are absolutely critical. Productive gathering may demand little or no coordination across people, but hunting large game requires more than one hunter working together to have any hope of success at all. At higher levels of technology, the point of Adam Smith's famous pin factory was to exemplify "increasing returns to scale" from coordination and specialization in production.[60]

[56] In the real world of hunter-gatherer groups, there are many tasks and an individual who is not strong enough to hunt large game can devote productive full-time effort to a range of other necessary work activities. Our simple model has only one productive activity. (Alternatively, you could think of $a_i e_i$ as total calories produced and a_i the best i can manage in any available activity.)

[57] In hunter-gatherer societies, skill definitely associates with prestige and status, and high prestige "big men" get considerable deference even as they are particularly inclined to share and display generosity. They also seem to be important for organizing cooperative endeavors. See Henrich, Chudek and Boyd (2015).

[58] Hint: In the utilitarian optimum case, the first-order condition gives you e_i^* as a function of $Y = \sum a_j e_j$. You can find equilibrium Y/n by multiplying e_i^* by a_i and then summing over all people to get an expression in terms of equilibrium Y. For the comparison of total output, you need to use Jensen's inequality: For a strictly concave function f, $f(\frac{1}{n}\sum x_i) > \frac{1}{n}\sum f(x_i)$.

[59] [Answers: $\bar{y}_{util} = (2b)^{-2/3}(\frac{1}{n}\sum_j a_j^2)^{2/3}$, and $\bar{y}_{indiv} = (2b)^{-2/3}\frac{1}{n}\sum_j a_i^{4/3}$.

[60] A common formalization of complementarity in production is the CES production function, which for our example would be $Y = (\sum_i (a_i e_i)^{-\rho})^{-1/\rho}$ where $\rho \geq -1$. ...

So it is interesting that even without any complementarities in production at all, the utilitarian social optimum demands greater total effort and produces more total output than if people were just working for themselves (provided there is variation in ability). Why is this?

With redistribution, the efforts of more productive types have a positive externality. They increase not only their own income but everyone else's as well, at a low cost in effort relative to lower-productivity types. As a result, collective wellbeing (in the utilitarian sense) is increased by having them work more than they would if producing only for themselves.[61] Net transfers flow from more productive types to less productive types.

In modern economies a person's skills and ability to produce high value output are to a great extent a function of education and training. These are themselves a function of public policy choices. ...

17.3 Insurance and the veil of ignorance

Hunter-gatherers did not read Jeremy Bentham or Karl Marx. How and why would they arrive at the utilitarian social optimum or, similarly, social arrangements that closely approximate "from each according to ability, to each according to need"?

Hunter-gatherers live close to subsistence and do not have good instruments for saving. An illness or injury that temporarily reduces one's ability to work would be catastrophic if everyone consumed only what they individually produced and there was no redistribution. When negative shocks to income can be fatal, a sharing norm – redistribution to equalize consumption – is a highly valuable form of *insurance*. Likewise, the anticipated shock of old age, when economic productivity declines, makes norms of intergenerational sharing attractive.[62]

The previous section showed that with concave preferences for material consumption, the naive utilitarian "first-best" arrangement involves redistribution to equalize consumption. This section shows how concave preferences plus uncertainty about individuals' ability to produce can make a sharing rule Pareto efficient, thus in the ex ante interests of everyone. No need for instruction in, or assent to, arguments from 19th century moral philosophy.

John Rawls (1971) and the literature on distributive justice focus a great deal on how to treat differences in "natural ability" and "social position" that are the result of luck, accidents of birth. Rawls famously proposes to deduce just social arrangements by asking what individuals behind "a veil of ignorance" – meaning prior to learning their a_i's, in the terms of the last section – would rationally agree to. He is thinking about a hypothetical deliberation.

But we awake every morning behind a veil of ignorance about what fortune has in store for us. And a great deal of actual politics concerns the ordering of political, social, and economic institutions to address the problems posed by random and intertemporal variation in our health and productive capabilities.

In the US, *by far* the largest categories of federal government expenditure are health and income support

[61] Again, at $e_i = 1$ this point is moot.

[62] Hill and Hurtado (2009) shows ...

for older citizens. Around X of government spending is on Medicare and Social Security, and another .... Attributed to Peter Fisher, the quip that the US government is basically an insurance company with an army is accurate and underappreciated. In addition to Social Security, Medicare, Medicaid, unemployment insurance, and many other income support programs, the federal government insures bank deposits, regulates risk in asset markets (SEC), and acts a lender of last resort to insure against financial crises. With the economic shock of the Covid pandemic, the US government provided 1.x trillion dollars in direct income support payments. Even defense spending is in part a kind of insurance.

Suppose that our group of n people decides on a transfer scheme behind a veil of ignorance about whether each one will be too sick to forage or hunt the next day. To make things transparent, suppose that they all have equal ability $a_i = a > 0$ if not sick, and $a_i = 0$ if sick. It would be natural to next suppose that each has an independent $\epsilon > 0$ chance of getting sick, but this makes for excess algebra since then total output is a random variable. Suppose instead that somehow it is known that exactly k people will be too sick to work, $0 < k < n$, and let $\epsilon = k/n$ be the probability that an individual is in the sick set.[63]

A transfer scheme that gives $t \geq 0$ to the k people who can't work costs kt. If each worker pays an equal share out of own production, this amounts to $kt/(n-k) = \epsilon t/(1-\epsilon)$, so behind the veil of ignorance an individual's expected wellbeing is

$$\epsilon u(t) + (1-\epsilon)u(ae_i - \epsilon t/(1-\epsilon)) - be_i^2/2.$$

Since everyone is the same behind the veil of ignorance, maximizing this individual payoff also maximizes the utilitarian sum. The first-order condition with respect to the insurance payment t equals zero when

$$\begin{aligned} \epsilon u'(t) &= (1-\epsilon)u'(ae_i - \epsilon t/(1-\epsilon))\frac{\epsilon}{1-\epsilon} \\ u'(t) &= u'(ae_i - \epsilon t/(1-\epsilon)). \end{aligned}$$

So, strict concavity of u implies that one wants to equalize consumption across uncertain states of the world – perfect insurance. The optimal t, given e_i, is $t^* = ae_i - \epsilon t^*/(1-\epsilon) = (1-\epsilon)ae_i$. Last, we find the optimal effort given t^* to be the e that solves $u'((1-\epsilon)ae) = be$. Sickness is effectively like a tax that makes effort less productive, which in turn makes leisure relatively more attractive.[64]

Exercise 24. *Analyze a more Rawlsian variant in which a continuum of "souls" will be assigned by natural lottery to have natural abilities $a_i \geq 0$ distributed by the smooth cdf F. Behind the veil of ignorance, they take preferences to be represented by the concave utility function for consumption $u(z) = z^\rho$, $\rho \in (0,1)$, and disutility for effort $be_i^2/2$. Assume that a_i and e_i will be observable and so they can design social arrangements to induce any effort level from any individual. Let $e(a)$ be the schedule that decrees effort as a function of observed ability, and for convenience allow any $e_i \geq 0$ (no upper bound).*

(a) Total output (and per capita income) will be $y = \int ae(a)f(a)da$. Assuming (correctly) that concave utility implies that ex ante individual utility will be maximized in a scheme that equalizes incomes after

[63] In a large society (large n) with independent sickness probability ϵ for each person, the share sick will be very close to ϵ.

[64] If the group is working very close to subsistence, sickness of some may mean that they may have to work harder when able; or more likely, $e^* = 1$ either way (corner solution) so this issue is moot.

transfers, write down the objective function behind the veil of ignorance, as a function of the work schedule $e(a)$.

(b) Derive optimal total output and $e(a)$ as a function of risk aversion, ρ, and the mean and variance of ability, $\bar{a}$ and σ_a^2.[65]

(c) Compare to total output and effort levels if each i independently chose e_i after observing own ability a_i. Ex post, after learning a_i, who is unhappy with veil-of-ignorance egalitarianism (if we assume that these utility functions are fully interpersonally comparable, not just in differences)?

(d) Rawls argued that souls behind the veil of ignorance would rationally choose social arrangements to maximize the wellbeing of the worst-off person (maximin, or the "difference principle"). If we take wellbeing to refer only to consumption and not disutility of effort, what is the implication in our model here?

(e) By contrast, what if we assume fully comparable utilities and ask what the difference principle would imply if we suppose that the objective behind the veil of ignorance is utility for consumption less disutility for effort? Finding the maximin effort schedule in this problem is mathematically difficult, but speculate about the (ironic) implications you would expect.[66]

What have we learned?

- The empirically accurate assumption of concave consumption preferences has strong implications for both desirable and actual political and economic institutions.
- At the individual level, concave preferences imply that people benefit from smoothing consumption over time and across uncertain states of the world, including sickness, disability, career luck, random economic shocks, and native ability before one gets a better sense of what this is. So there is a strong welfarist argument for redistribution from the fortunate to the unfortunate as good policy.
- Because this implication follows from an ex ante common interest across persons (ex ante value for insurance), it does not depend on a stronger assumption of utilitarianism for justification. However, with concave consumption preferences, the naive utilitarian standard (policy should maximize the sum of a comparable measure of individual wellbeing) implies that good policy would ideally redistribute to equalize consumption, veil of ignorance or not.[67]
- Given an ex ante efficient insurance scheme, ex post luckier, more privileged, and higher income types would provide net transfers to less privileged, less lucky, and lower income types. In

[65] You can find $e(a)$ by "pointwise maximization"; basically, look at how it worked above in the discrete case.

[66] Alternatively, solve the problem when there are two types, $a_h > a_l > 0$. Hint: Argue that the maximin must equalize their ex post utilities.

[67] As is often noted in political philosophy treatments, the utilitarian principle of maximizing a sum of measures of wellbeing is egalitarian in the sense that wellbeing is weighted equally across people, but at the same time is completely uninterested in equalizing happiness levels across people. Its core moral principle does not know about fairness. If $(20, 20, 20)$ and $(0, 0, 61)$ are two distributions of comparable utilities across three people, naive utilitarianism says the second is better, and it is indifferent between $(50, 50)$ and $(0, 100)$. With concave utility over consumption, it just happens mathematically that maximizing the sum requires an egalitarian distribution, observed in this context by Edgeworth (1897).

a large, modern society, it is easy to see that *the well-off may not like this* and would have a preference for lower tax rates and transfers. So the analysis suggests a basis for the left-right economic dimension that has tended to dominate politics in modern economies (more below).

- So far we have been asking about socially optimal redistribution under very strong assumptions about ability to observe people's abilities a_i and to monitor work effort e_i. While this is reasonable for small groups of hunter-gatherers, it is not at all reasonable for large, technologically complex modern economies – or, in fact, for important features of agricultural societies (see, for instance, share-cropping).

17.4 Optimal tax policy

Recall our initial formulation in which individual i's preferences are described by $u(a_i e_i - t_i) - be_i^2/2$ and we want to identify the tax and transfer policy $t = (t_1, t_2, \ldots, t_n)$, $\sum_i t_i = 0$, that maximizes the objective

$$\sum_i u(a_i e_i - t_i) - be_i^2/2.$$

We saw that if we could condition transfers on individual effort and ability levels, utilitarian redistribution would equalize consumption across people, implying individual payoffs of $u(Y/n) - be_i^2$ where $Y = \sum a_i e_i$ is total output. And similarly, if we choose taxes and incentives behind a veil of ignorance about a_i, the optimal insurance scheme would equalize after-tax incomes.

Suppose now that effort and ability are *not* observable, and that individuals are guaranteed Y/n and find work unpleasant relative to not work. Then in a large population – where one's output is going to be spread over many people and provides almost no direct benefit to self – clearly everyone in this model would choose $e_i^* \approx 0$, so there would be basically nothing produced. This very bad outcome is the result of a moral hazard problem. People would like to collectively commit to positive effort levels but have a private incentive to slack off.[68]

Mirrlees (1971) posed and analyzed a somewhat more general version of this optimal tax problem under the assumption that neither effort nor ability is observable. The social planner knows the distribution of the a_i's and chooses a tax schedule $t(y)$, which is the tax paid or, if negative, transfer received, by a person whose income is observed to be y. An efficiency-equity tradeoff is then immediate. People need to be able to keep some of their own earnings to have an incentive to be productive, but disparities in a_i's will then lead to disparities in after-tax income. Even if we are designing a tax scheme behind the veil of ignorance for the purpose of insuring people against random natural ability, career luck, and unequal inheritances, the optimal scheme must involve less than perfect insurance.

Mirrlees' model is mathematically difficult. Even if we restrict attention to a linear tax rate $r \in [0, 1]$ with a "demogrant" $b \geq 0$ ($t_i = ry_i + b$), it is hard to find functional-form assumptions that allow an explicit solution for the best r and b. Piketty and Saez (2013) provide a general analysis of this case, which they note is of particular interest because empirically "To a first approximation, the tax burden is

[68] With our simple deterministic production technology, $y_i = a_i e_i$, if we can observe one of either ability or effort we can infer the other. More realistically, there are unobservable random influences on output, or measurement error (see, for example, Academics), so that inability to observe either one creates problems.

distributed proportionally to income" for OECD countries (395). The qualitative features of an optimal linear income tax are straightforward. Rather obviously, the optimal linear tax rate (with the naive utilitarian standard) should be lower the more sensitive effort is to tax rates, and it should be higher the greater the degree of inequality in a_i.[69]

[non-linear optimal income taxes]

17.5 Paying for public goods

So far the only function of government we have considered is redistribution for purposes of equity or insurance. What about government spending on so-called *public goods* like roads, water, sanitation, and power infrastructure, public schools, national defense, environmental and health and safety regulation, and the judicial system?

Our basic approach is easily adapted by having taxes go to a good that all taxpayers like, if possibly in different amounts. What will the socially optimal amount be, and who will want more or less?

The following set up is useful and often used in basic political economy models. We represent society as a continuum of individuals (or households), $i \in [0, 1]$, where the pre-tax income of individual i is $y_i \geq 0$. The distribution of pre-tax income in society is described by the cdf $F(y)$, with mean and median income being $\bar{y}$ and y_m, respectively. (So we put aside the question of incentives to work and just take pre-tax income as given.) In all known empirical cases, income distribution is significantly skewed right, so that the average income $\bar{y}$ is greater than median income y_m. In the US, 2019 median household income was about $69,000 versus $117,000 for average income.

A proportional (or linear) tax $t \in [0, 1]$ funds an aggregate public good $g \geq 0$, so that $g = t \int y_i dF = t\bar{y}$. Because we are representing society as a continuum with unit mass, g is both total and per capita public good spending.

Suppose that people like both more consumption out of after-tax income, $c_i = (1 - t)y_i$, and more public goods, g. We represent preferences with an interpersonally comparable utility function $u(c_i, g)$, increasing in both arguments.

Exercise 25. *Consider the specific example $u(c_i, g) = 2\sqrt{c_i} + 2a\sqrt{g}$, where $a > 0$ scales the value of public goods relative to consumption (the 2's just help clean up the results a bit). Solve for an individual's most preferred tax rate $t^*(y_i)$, and discuss how and why this varies with pre-tax income y_i and with average income $\bar{y}$.*

Points:

- Higher-income types prefer lower tax rates, just as in our analysis above for higher a_i. The reason is slightly different from the straight redistribution models. Here, everyone is getting the same value from public goods at level g, but the price tag for these goods differs by pre-tax income.

[69] Less can be said in general about the structure of optimal non-linear tax schedules, $t(y)$, except that the marginal tax should decline towards zero at the very top end (i.e., should approach a flat tax above some level).

It is ty_i, so higher y_i means you are paying more for the same bundle of goods, for instance, national defense and well-maintained roads. With these preferences, the higher price tag means that higher-income types demand less and smaller "government."

So again we have a plausible proposal for why advanced industrial economies that pay for government services with an income tax tend to have political conflict organized along a Left-Right cleavage, with higher-income households wanting lower taxes, spending, and redistribution, and lower-income households preferring the reverse.

- When funded by a proportional or progressive tax, there is a sense in which public goods are a kind of redistribution. Public goods effectively redistribute some overall wellbeing from unequal private consumption to more equal public consumption and investment (for instance, via public parks or publicly funded education).

The baseline Downsian model of electoral competition will be explored ad nauseam in subsequent sections. You will see that with two parties choosing platforms and voters voting for the party whose platform is closer to their favorite policy, the parties converge to the policy most preferred by the median voter.

So if the parties or candidates are proposing tax rates, the tax rate preferred by the median voter would be proposed in equilibrium by both parties. This implies a higher tax rate and a larger government sector than the voter with the average income would prefer. Further, if we fix per capita income $\bar{y}$ and increase income inequality, y_m decreases, implying pressure for higher tax rates and more redistribution via public goods.

This point and the model behind it is associated with Meltzer and Richard (1981). It is heavily employed in work by Acemoglu and Robinson (2006) on the political economy of democratization, where it motivates class conflict over whether rich elites or masses (via democracy) control government.[70]

But what does the socially optimal tax rate/size of government look like, using our naive utilitarian or social-peace standard? The next exercise asks you to work this out.

Exercise 26. *Find the tax rate that maximizes the sum (or average) of $u(c_i, g) = 2\sqrt{c_i} + 2a\sqrt{g}$, in a society with n people and incomes y_i. Interpret. How would greater variance in y_i (income inequality) affect the optimal tax rate? Why is this? How does the optimal tax rate compare to the tax rate preferred by the person with average income $\bar{y}$?*

Income effects on demand for public goods. Higher-income households pay more for the same amount of public goods, but they also have more income they can use to pay! If government services are a "normal good," meaning that people want more as their incomes rise, then this "income effect" works in the other direction from the "substitution effect" described above. Put differently, while low-income households get more public goods per tax dollar paid, they also have lower incomes and so might be more sensitive to taxes relative to other uses for their disposable income.

The preferences we used above, $u(c_i, g) = 2\sqrt{c_i} + 2a\sqrt{g}$, happen to be special in a way that implies that changing income levels without changing relative prices does not affect the proportions of c and g that

[70] The basic idea also appeared in Aristotle ...

a person wants.[71] Let's consider the more general formulation $u(c_i, g)$, increasing in both arguments at a decreasing rate.

For household i and tax rates $t \in [0, 1]$, only certain combinations of consumption and public goods are feasible. Together, $c_i = y_i(1 - t)$ and $g = t\bar{y}$ imply $c_i = y_i(1 - g/\bar{y})$ or

$$y_i = c_i + \frac{y_i}{\bar{y}} g.$$

This is like a budget constraint, with y_i being total income (budget) but also increasing the price of g relative to c. So you can see directly that higher income gives you more to "spend" on both consumption and government services, but at the same time makes g relatively more expensive. Greater per capita income $\bar{y}$ relative to y_i makes government spending more attractive, because with a larger tax base ($\bar{y}$) the same tax rate generates more spending.

Exercise 27. *Solve the first-order condition for the household's problem* $\max_{c_i,g} u(c_i, g)$ *subject to the budget constraint, and express in terms of the implicitly defined optimal tax rate.*

Figure 9 illustrates using the standard diagram that shows the household's optimal (c_i, g) mix where its indifference curve from $u(c_i, g)$ is tangent to the budget constraint. As you found, this is where the marginal rate of substitution in utility is the same as the relative price of c versus g. Tax rates vary from 1 to zero along the budget constraint (moving northwest to southeast).

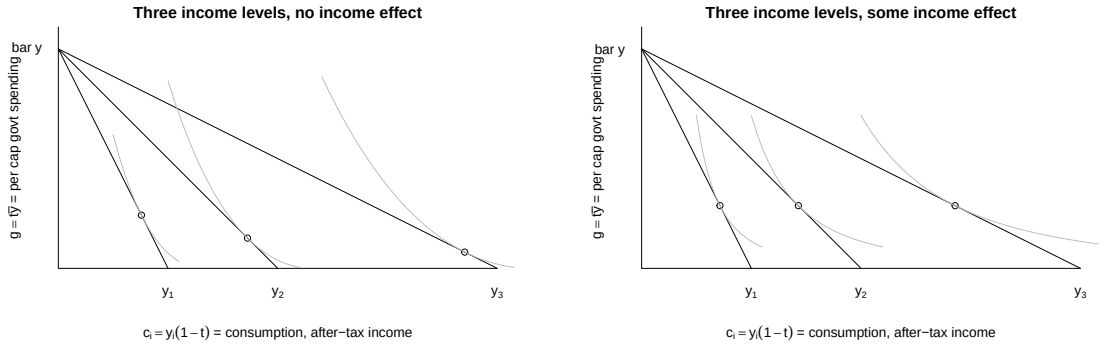
The left panel shows the case of $u(c_i, g) = 2\sqrt{c_i} + 2a\sqrt{g}$, where there is no income effect and preferred tax rate/size of government decreases with household income. The right panel shows what happens with preferences $u(c_i, g) = \log c_i + a \log g$, which is a special case in which the income and substitution effects of increasing y_i exactly offset each other. There, you see that a household's most preferred tax rate/size of government does not vary with household income. This is because while higher-income households pay a higher tax price for public goods, higher income also makes them want to "buy" more.

For public goods where rich people can more readily substitute with private good alternatives – for instance, old-age and health insurance, private versus public schools, or vacations to Hawai'i rather than time in the public park – we would expect smaller income effects and thus more "class conflict." For other public goods – like breathable air, clean water, and perhaps national defense – it is harder to purchase private substitutes, and we might expect relatively more support for spending by higher-income types.

Acemoglu and Robinson (2000) argued that successive expansions of the franchise in 19th century Britain lowered the income of the median voter, leading to greater redistributive public goods spending. Lizzeri and Persico (2004) suggest that instead, much of the increased spending was on health and sanitation in the cities, reflecting increased demand by urban elites for a public good (public health in rapidly growing cities) that they could not readily substitute for in the face of diseases like cholera.

[71] Formally, this u is *homothetic*, meaning that it can be expressed as a monotone transformation of a function that is homogeneous of degree 1. The relevant consequence of these features is that the marginal rate of substitution between the goods c and g depends on the ratio of c to g but not the levels.

Figure 9: Preferences for public goods/size of gov't vs disposable income



17.6 Taxation in the "natural state"

Even if both are institutions that address a universal human desire for insurance, sharing in hunter-gatherer groups is one thing, and the welfare state is quite another. In between there is a vast technological and institutional terrain, almost none of which involves income taxes.

With occasional exceptions, income taxes have appeared only recently in human history, as they require high levels of state bureaucratic and technical capacity to collect. The British government used them briefly during the wars following the French Revolution. They then start to appear gradually in more and more countries from the late 19th century on. In the early 20th century levels in OECD countries were around 10% of GDP, rising to around 40% on average today. Piketty and Saez (2013) observe that

> the historical rise in the tax burden has been made possible by the ability of the government to monitor income flows in the modern economy and hence impose payroll taxes, profits taxes, income taxes, and value-added-taxes, based on the corresponding income and consumption flows. Before the 20th century, the government was largely limited to property and presumptive taxes, and taxes on a few specific goods for which transactions were observable. Such archaic taxes severely limited the tax capacity of the government and tax to national income ratios were low (see Ardant, 1971 and Webber & Wildavsky, 1986 for a detailed history of taxation).[72]

In addition to property taxes, for centuries a common method of raising revenue has been *the mafia model*, termed "the natural state" by North et al. (2009). Appearing in multiple forms in both contemporary autocracies and democracies, this system is anything but "archaic" and it is plausibly a major

[72]The US government enlists employers in the collection of income and other taxes through the system of withholding from paychecks. ...

part of the explanation for why things are often so messed up in terms of government performance.

The mafioso, a "specialist in violence," visits a business owner with a proposal: Share your revenues with me and I will protect you from other specialists in violence. Implicitly, "other" specialists in violence include his own organization. "It would be a shame if something terrible happened to your business." The mafioso is a "stationary bandit" who uses the threat of force to extract from producers in his zone, but also has incentives to protect them from rival predators and ordinary crime; to invest to some degree in public goods that will make them more productive; and to not extract so much that their avoidance and effort choices reduce his total take (Gambetta, 1993; Olson, 1993; Sánchez de la Sierra, 2020).

Likewise, the king in 17th century England and the modern regimes that practice "crony capitalism" make deals with elite business partners, using the state's coercive capabilities to protect them from economic competition. This confers monopoly or oligopoly rents on the cronies, which are then shared with or otherwise extracted by the ruler. Instead of having to monitor and collect from thousands of low-income households, the regime gets both big-money revenue generation with low administrative overhead, *and* a base of rich political supporters who are at risk of losing their oligopoly rents if the ruler loses power. At the same time, non-elites appreciate that the rulers are not trying all that hard to tax their household income or transactions, even as they bemoan the high prices, low employment, inequality, and inefficient and corrupt services in an economy shot through by state-sanctioned oligopolies.

This "natural state"/mafia/crony-capitalist arrangement makes so much political sense that it exerts a kind of gravitational pull for any and all political systems above some level of economic development, democratic or not. Although there can be interesting exceptions, the system is generally not good for economic growth.

Why is that? In the first place, as already suggested, state-sanctioned and enforced monopolies and oligopolies prevent competition and entrants that would tend to lower prices, increase employment, and increase quality of services. Second, incentives for capital investment and innovation to increase productivity – the driver of economic growth – are typically weak in these systems. Suppose a business owner in the mafioso's jurisdiction could greatly increase revenues and employment by making a capital investment in new and more technologically advanced equipment. It is in the mafioso's interest that the entrepreneur do this. But the entrepreneur anticipates that the mafioso, or the ruling family, will inevitably tax away a share of the gains. This lowers the incentive to invest in the first place. The mafioso may try to commit himself to a low tax rate to encourage investment, but are such promises credible? We will return to dynamic models of this kind of commitment problem in the last part of the book.[73]

Even if the mafioso ruler can commit to limit expropriation of gains from investment and development, he often has another powerful reason to stick with crony capitalism: Fear of rich rivals who are not beholden to him. Wealth can be converted to political power and influence. This is another commitment problem – this time on the side of other oligarchs or potential opposition to a ruling faction – that we will consider in section X when analyzing the problem of power sharing.[74]

[73] Note that the underlying tradeoff is the same as in the models of optimal taxation – higher taxes, less effort or investment. But in the natural state the purpose is redistribution to the ruler rather than social insurance.

[74] An argument to this effect concerning China that appeared just as I wrote this: "Mr. Xi is engaged in a systematic campaign to remove or neutralize people who have amassed a fortune. . . . This campaign threatens to destroy the geese that

The point for now is that taxation in the mafia system of the "natural state" is organized in a manner that is administratively low cost; relatively invisible (or disguised) to most people; serves a ruling faction's interest in maintaining power; but is economically dysfunctional. Societal *coordination* on the political, administrative, and technical capability to monitor and collect taxes on individual income and transactions is no simple feat. For example, how to convince large numbers of workers and business owners to pay taxes when they believe, with much prior evidence, that the state is a criminal, predatory gang?

18 Why electoral democracy?

So far, we have implicitly posed the problem of good government as how to choose a large set of policies so as to maximize a social welfare function that reflects some combination of naive-utilitarian and social-peace objectives. In both approaches, government performance is better, the better the match between policies and a weighted average of individual preferences.

If that's the problem, why use elections to select people to figure out and implement policies?

You might ask, Why not? and What's the alternative? If we want a correspondence between popular preferences and policy, and if popular preferences may change and shift around, why *not* use a method that selects policymakers from the population? Candidates and parties offer policy positions in electoral competition, and election results then reveal what "the people" want at that time.

Part III considers a series of models that are based on assumptions that should be highly favorable to this argument for electoral democracy. In these "Downsian" models, voters are assumed to know both what policies they want and what policies politicians will implement if they win office in the election. We will see that, surprisingly, even on these favorable assumptions it does not follow *at all* that elections will produce policies close to what the average or even median voter wants. In a nutshell, it is difficult to devise an electoral and constitutional system that incentivizes and selects only political moderates.

This is not an abstract, hypothetical problem. In the US, for example, Democratic and Republican candidates in House races have repeatedly been shown to take policy positions much closer to – or even more extreme than – the typical preferences of their co-partisans rather than the average or median voter in their district.

Is there an alternative that would work better to match policies to public preferences?[75] How about *sortition* – choose political representatives at random, by lottery. If you are concerned that we want

lay the golden eggs . . . Now private and state-owned companies are being run not only by their management but also a party representative who ranks higher than the company president. This creates a perverse incentive not to innovate but to await instructions from higher authorities." George Soros, "Xi's Dictatorship Threatens the Chinese State," *Wall Street Journal*, 14-15 August 2021, A11. "Oligarchs" in Putin's Russia also know a thing or two about the dangers of the ruler thinking that one might pose a political threat.

[75] By "preferences" I am referring not just to what people say they want in public opinion surveys, which is not necessarily what they would want if well-informed. Also, policy questions are selected for public opinion surveys on the basis of what politicians choose to fight over and what the media chooses to emphasize. One consequence is that the surveys that political scientists study rarely ask about perceptions of the quality of basic government services and outputs. "Preferences" here can and should refer to preferences over outcomes people care about (e.g., full employment, good health insurance, and so on).

competent, well-qualified people, fine: Restrict the lottery to candidates who meet some set of requirements, as well as being willing to serve. If you are concerned that this system is too "high variance," in that one random person's political views could be all over the map, then have multiple randomly chosen representatives for each district, and send them to a national legislature large enough that the average member would be close to the national average citizen. Represent the population with a random sample![76]

Such a political system is clearly a democracy but it is not an electoral democracy. The thought experiment is useful for helping to identify the value and purpose of using elections.

With a "lottocracy" as described, representatives have no ability to develop a career as a political leader, and thus no incentive to act to promote public interests while in office besides honor, personal integrity, and threat of legal sanction from being found guilty of corrupt behavior. Such a system would necessarily invest enormous power in the permanent bureaucracy and courts, whose nominal bosses come and go and who have essentially no understanding of how things work, in terms of policy and policy making. The nominal bosses would be at great risk of developing good means to profit at public expense in their brief time in office (having literally "won the lottery"), probably in tacit and explicit collusion with members of the permanent bureaucracy. The citizens would have no recourse, no way to sanction poor performance if the system devolved into a predatory elite that essentially pays off a new set of lucky lottery winners every few years.

The point is that while different electoral and constitutional designs can be better or worse at selecting moderate and capable politicians – an important subject for political science; see below – the central justification for electoral democracy lies in the need to give citizens the ability to coordinate to remove power-holders. We use elections to try to protect against extreme outcomes like autocracy.[77]

To some degree, elections may also serve to aggregate and select policies (via politicians) that are close to typical citizen preferences; to select capable and qualified leaders; and to give politicians a reason to perform well lest they lose their next election. But as we will see – and as we experience in everyday life – elections by themselves should not be expected to be perfectly effective for these goals.

19 Downsian competition in a policy space

Section 14.2 analyzed a model with two politicians choosing policy positions in the shadow of an election. This kind of model – in which candidates or parties choose issue positions in a (typically one-dimensional) issue space, and then voters vote based at least partly on their platforms – are often called Downsian after the analysis of Anthony Downs (1957).

[76] In fact, the same problems of preference aggregation discussed below for electoral systems to select representatives apply within a legislature for voting on bills, so random selection to a legislature does not entirely avoid these difficulties unless it operates by randomly selecting a member to decide each matter. Which would have other problems.

[77] See Landemore (2020) for an extended and interesting set of arguments for lottocracy. She allows that lack of accountability might cause problems, but counterpunches, suggesting that elections may not be much better on this score. I agree that elections are a blunt and unwieldy instrument for mitigating politicians' moral hazard, but they are far from useless and also can provide an essential if imperfect check on very bad outcomes associated with dictatorship (Fearon, 1999, 2011). The subject of moral hazard in electoral control comes up repeatedly in the sections on Downsian models that follow, and is addressed more directly in the sections on retrospective voting models.

Part IV presents and analyzes a series of variations and extensions of the baseline Downsian model. Our overarching purpose there will be to develop a better understanding of reasons for democratic failure, in the sense of failure of electoral democracy to bring about policies that are good for most citizens.

In most, but not quite all, of the Downsian models that we consider, such failures can arise *even though* voters are assumed to be able to observe and condition their votes on what politicians are actually doing. We explore arguments based on weakening this obviously strong assumption in section 41, under the heading of "moral hazard" models of electoral control.

It is striking, however, that there can be good reasons to expect substantial misalignment of public policies with voter preferences even when voters basically "know what's going on." I do not know whether democratic failure due to moral hazard problems (the difficulty of monitoring and evaluating what politicians are doing) or the reasons for non-moderate policies in Downsian models are more important for welfare. But I think there is ample empirical evidence that the latter are quite significant.

Before we get into the (somewhat) more positive and explanatory analyses of Part IV, however, we need to do more normative work. ...

19.1 Preferences on a policy dimension

To begin, we consider a less clunky, more standard version of the candidate-platform choice game of section 14.2. We represent the electorate as a *continuum* of voter *ideal points* distributed on some interval of the real line. There is no harm in using the interval $X = [0, 1]$, and imagining that $x \in [0, 1]$ represents a policy position on the left-right spectrum. Depending on the political situation we are thinking about, we could imagine this interval as standing for

- positions on a particular issue – for instance, possible public policies on abortion, ranging from "on demand, with insurance coverage" to "never under any circumstances."
- Or in some models we will take the interval literally, as representing possible tax rates from zero to 1.
- Or we can imagine the interval to be summarizing a general or aggregated left-right dimension. This is often reasonable in the US political context at least, given how strongly legislators' policy preferences on a large range of issues scale to approximately one dimension.
- Or it could represent degree of attachment to or identification with an ethnic or political category, like party identification. The dimension does not have to be a literal representation of policy view or preferences at all. As the field of Political Behavior emphasizes, many voters appear to derive some of their expressed policy preferences from identification with a party, rather than the other way around. As we will see (section 41), this can make a lot of sense if monitoring and evaluation of what politicians do in office is difficult or impossible. It can then be reasonable for voters to vote based on judgements of how similar a candidate is to them as a person, especially as regards views on politics and society. In this reading, ideal points on $[0, 1]$ could represent a range from very Left to very Right social types, or rather policy stances associated with these types. But we

understand that when it comes to voter decision-making, many are using social and party cues rather than distinct information and knowledge about policies.

Let $F(x)$ model the distribution of voter ideal points, meaning that $F(x)$ is the share of the population whose ideal points are less than or equal to $x \in [0, 1]$. With our bounded interval and assuming a smooth distribution, $F(0) = 0$, $F(1) = 1$, and of course F is increasing on the interval.

Let x_i be voter i's *ideal point*. This means that x_i is i's most preferred policy in $[0, 1]$: $x_i \succ_i x'$ for all $x' \neq x_i$. In most applications, it is reasonable to assume further that individuals' preferences on the issue space are *single-peaked*, in the sense that outcomes further in one direction from the ideal point are considered less good. Formally, i's preferences on X are single-peaked if $x_i \succ_i x' \succ_i x''$ whenever either $x'' < x' < x_i$ or $x_i < x' < x''$.

In part because we usually are not attaching any objective metric to X, it is convenient to assume further that voters' preferences are symmetric in $[0, 1]$, so they prefer outcomes that are closer to their ideal point rather than farther away. That is, for policy positions $x', x'' \in [0, 1]$, $x' \succeq_i x''$ iff $|x_i - x'| \leq |x'' - x_i|$.

In what follows we will constantly use two specific representations of symmetric, single-peaked preferences, termed "linear" and "quadratic." Illustrated in Figure 10, these are, respectively,

$$u_i(x) = -a_i|x - x_i| \text{ and} \tag{3}$$

$$u_i(x) = -a_i(x - x_i)^2, \tag{4}$$

where $a_i > 0$ scales i's "utility loss" with respect to policy departures from x_i in the metric attached to the policy space X.

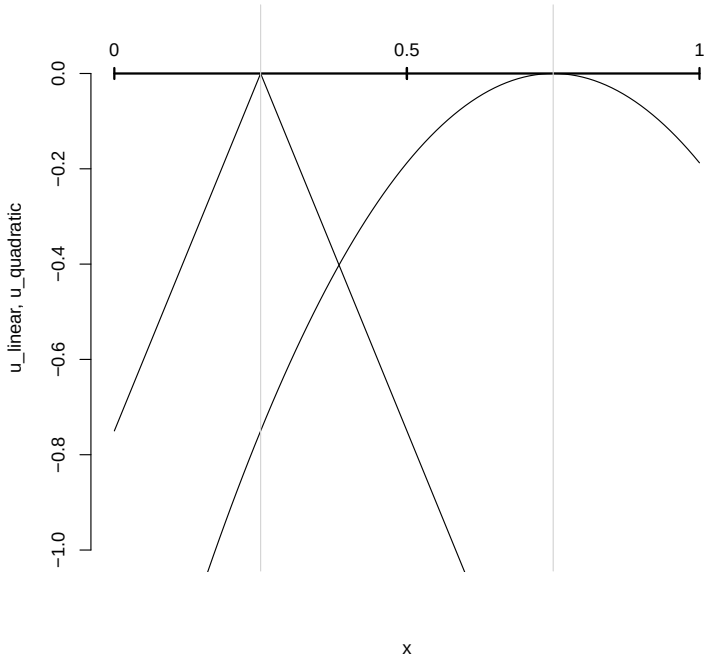
You should be able to convince yourself that these payoff functions represent exactly the same ordinal preferences for voter i. With expected utility, they no longer represent quite the same preferences. The difference is subtle, and often does not matter when it comes to implications for candidates' equilibrium actions. But when it comes to *normative* implications for voter wellbeing, there is a substantively meaningful and important difference.

In both cases voters are modelled as *risk averse* concerning lotteries that might give them outcomes on either side of their ideal point. The difference is that with the quadratic-loss preferences, the voter is also risk averse concerning lotteries on outcomes that are all on one side of her ideal point.

Substantively this means that with quadratic preferences the voter's alarm and unhappiness grow at an increasing rate as policy/candidates move away from her ideal point. For example, consider a moderate voter contemplating the difference between a center-right and a hard-right candidate or policy position (Romney and Cruz?). Suppose that she has linear preferences and happens to be indifferent between the center-right position and a 50-50 lottery on (moderate, hard right). If she had quadratic preferences instead, then she would strictly prefer the center-right position, because with quadratic preferences she sees hard-right as (relatively) *really bad*.

In Section 21 we will consider in greater detail the normative implications of quadratic versus linear preferences for what is good policy. To preview, with quadratic preferences people with more extreme (away from the center) policy preferences are especially unhappy, which is bad both from a utilitarian

Figure 10: Example single-peaked linear and quadratic preferences



perspective and from a social-peace perspective. The implication is that for a skewed distribution of voter ideal points, with quadratic preferences the best policy should arguably be closer to the *mean* rather than the median, because the mean minimizes the average squared distance (quadratic loss).

Figure 11 illustrates several distributions of voter ideal points, along with vertical bars showing the mean and median positions. Figures 11(a) and (b) are smooths of draws from, respectively, a symmetric unimodal distribution and a symmetric bimodal distribution.

Figure 11(c) is a smooth of the actual 2012 Republican presidential vote share in the 435 House districts. This is a good first approximation for average or median ideological position of voters in a district; and since districts are all the same size, it is not bad as an approximation for the distribution of voter ideology across the country.[78] Note that some left skew opens up a small gap between mean and median in the population.

Figure 11(d) is a smooth of Adam Bonica's estimates of the Left-Right ideological position of members of the US House elected in 2012, based on campaign contribution data.[79] Representatives' Left-Right ideologies are strongly bimodal, with the median member much more conservative than the average representative. (Republicans won 54% of House seats in 2012.)

[78] Rodden (2019, 201) shows a graph of district ideology based on a much more sophisticated method of using individual survey data, based on Tausanovitch and Warshaw (2013) and his work with McCarty et al. (2019). It is extremely similar.

[79] (Bonica, 2018). I rescale the CF scores to [0, 1]. The picture looks almost the same if one uses DW-Nominate scores.

In brief, if the measures of representatives' left-right preferences are accurate, then the difference between 11(c) and 11(d) means that there are many congressional districts with moderate average voters who are being represented by politicians who are much less moderate in their expressed policy positions.[80]

A last point about these spatial representations of policy preferences: It is both highly convenient and often intuitively compelling to posit a continuous dimension on which policy positions can be located. In some cases, like for tax rates, there is an objective continuous metric. In other cases, like policies on abortion or gun control, it is not too difficult to locate discrete policy descriptions in an ordering from, say, least restrictive to most restrictive. Further, political scientists have developed sophisticated methods of using data on roll-call votes in Congress, campaign contributions, and survey responses on policy questions to scale politicians' and voters' policy preferences to points on one or more continuous dimensions. In order to produce the estimates, these methods assume the existence of latent, continuous policy dimensions, and also assume that actor preferences over these latent dimensions take a particular form, typically quadratic utility.

But we should not take these representations too literally, especially when there is no objective metric defining positions on a policy dimension. In formal terms, Eguia (2013) identifies the conditions needed for single-peaked preferences over a multi-attribute policy space to be representable by concave utility functions on an abstract Euclidean policy space. It is hard to say, but they may be demanding. More on the structure of policy preferences in sections XXX below.

19.2 The baseline Downsian model

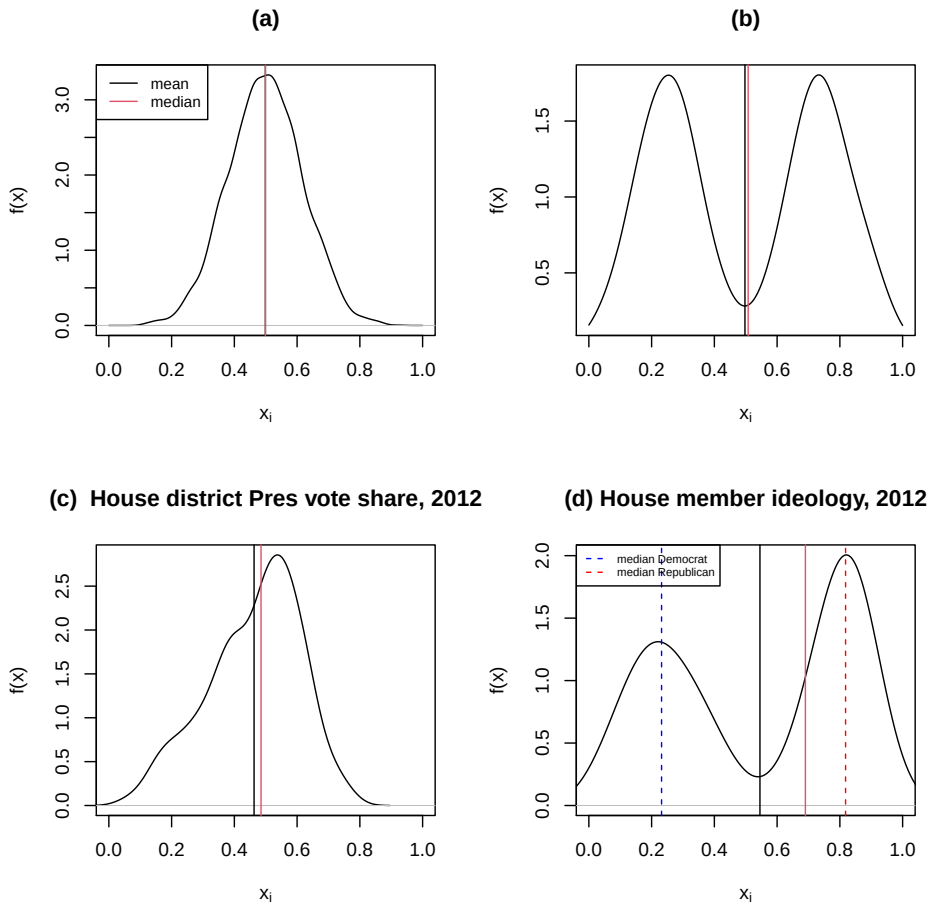
The baseline Downsian model is as follows. We consider the two-player normal-form game in which two candidates, L and R, simultaneously choose platforms, call them l and r. Voters vote for the candidate whose platform is closest to their ideal point, and they split the total vote equally if they choose the same platform. Thus, if it happens that $l < r$, L gets $F((l + r)/2)$ share of the vote, and R gets 1 minus this. If $l = r$, they each get half the vote.

What to assume about the candidates' objectives? The two standard choices for the baseline Downsian model are (a) they try to maximize vote share, and (b) they try to maximize the probability of winning a majority. As you will see, in this model it does not matter which we assume. However, in both cases some annoying technical issues arise.

With vote share as the objective, the technical issue is that with our convenient depiction of the issue space as a continuum, best replies do not generally exist. For example, if the median is at .5 and $r = .6$, Left wants to choose the position closest to .6 but less than .6, and there is no such point. In what follows, here and in general, we can finesse this issue when it arises by assuming that we are really talking about an issue space with a large number of discrete positions, for instance $X = [0, 1/n, 2/n, \ldots, 1 - 1/n, 1]$, where $n > 0$ is arbitrarily large. It is not realistic that people can identify or

[80] As discussed in section X below, there are reasons to think that standard estimates of US politicians' positions may make them look more extreme than they really are, relative to median or mean opinion in their states or districts. It is still clearly the case, however, that Republican and Democratic politicians represent the same states and districts very differently, which is prima facie evidence of significant divergence from median voter preferences.

Figure 11: Example distributions of voter ideal points



distinguish between fine-enough variations anyway, so this is no matter. The continuum is an analytical convenience.[81]

Exercise 28. *(a) Find any and all pure-strategy Nash equilibria of this game, under the assumption that candidates want to maximize votes. Hints: Do not try to draw a matrix. Instead, a good approach is often to make a guess, and then check whether it is in fact Nash. That is, ask yourself if any player has an incentive to deviate to a different action given what the other is choosing. If your guess was not a Nash equilibrium, try again. As you repeat this process, you should start to gain a better understanding of the logic of the strategic situation.*

(b) Repeat, but assuming that candidates try to maximize their probability of winning.

(c) How does the distribution of voter preferences affect the policy outcome in this simple model? Consider as examples the distributions shown in Figure 11. In your opinion, is this a good outcome or a bad outcome? Is there an outcome you think would be better for the voters as a society? If so, why?

19.3 Policy choice versus consumer choice

Next let's consider a classic example from the theory of markets, introduced by Harold Hotelling (1929). Hotelling suggested that the basic logic applied to candidates or parties choosing positions, and he thought this a bad thing:

> So general is this tendency [for "excessive sameness"] that it appears in the most diverse fields of competitive activity, even quite apart from what is called economic life. In politics it is strikingly exemplified. The competition for votes between the Republican and Democratic parties does not lead to a clear drawing of issues, an adoption of two strongly contrasted positions between which the voter may choose. Instead, each party strives to make its platform as much like the other's as possible.

Imagine a long length of beach by the ocean that we represent with the unit interval, $[0, 1]$. The city gives permits to two ice-cream vendors, who are allowed to position themselves wherever they want along the beach. Beachgoers uniformly distribute themselves along the beach. Each beachgoer has unit demand for an ice cream cone (they each want 1), and pay a cost of $c < 1$ per distance travelled to get to the seller. So they always go to the closest seller (the products are the same, and assume they charge the same price).

If we approximate the beachgoers with a uniform distribution on $[0, 1]$ and the sellers locate at x and y in $[0, 1]$ with $x < y$, then the x seller gets share $(x + y)/2$ of the business and the seller at y gets $1 - (x + y)/2$. If they locate at the same spot, they split the market equally.

[81] If we assume probability of winning as the candidates' objective, then best replies always exist but there can be lots of them. This is not really a problem in any meaningful sense, as you will see in the exercise. These are both examples of "technical artifacts."

Here is an alternative, more political interpretation of this model: $[0, 1]$ is a continuum of media consumers, who choose which television news network to listen to. $x_i \in [0, 1]$ represents the left-right political orientation of listener i, so that $x_i = 0$ is an extreme left-winger and $x_i = 1$ is an extreme right-winger. Suppose that individuals prefer to listen to the network whose political slant is closer to their own political preferences. For the game, the TV networks simultaneously choose their slant, or bias, $b_j \in [0, 1]$ for networks $j = 1, 2, \ldots, n$. We could also model entry – for instance, suppose there is a fixed cost $\delta > 0$ to enter the market, and the network's payoff is its market share.

Exercise 29. *For the Hotelling game with two players (the beach example):*

(a) Using the above payoff functions, find the unique Nash equilibrium of the location game between the sellers.

(b) Assuming that a customer gets value of 1 for the ice cream cone, compute the average payoff for beachgoers in the Nash equilibrium.

(c) In terms of beachgoer welfare, what would be the socially optimal placement of the two ice-cream vendors?

(d) Compare this situation to the Downsian model of candidate platform choice above. What is different about the two situations that bears on what is (arguably) the best policy outcome for society?

In the political model, our implicit assumption is that there has to be *one policy or representative* for all voters in the constituency, which could be congressional district or an entire country. The idea is that we are trying to choose – or for some reason we are constrained to choose – a *public* policy that applies to everyone. By contrast, in the market location model it is perfectly possible to have two different locations for the vendors. Hotelling's is a model of *product differentiation*. Often or typically, consumer welfare is improved by having a greater variety of goods and services available.

Perhaps there are reasons to think it is a bad thing if political parties converge to the median voter's preferred policy – much more on this question below. But this is not clear, because there is a fundamental *dis*analogy between the policy and market settings that undermines Hotelling's argument on this score, at least concerning a single public policy for single jurisdiction.[82]

To drive the point home: If the market supplies two types of cereal, Rice Krispies and Corn Flakes, this is almost surely better than if it supplied only one. People can choose which one they like better, and my consumption of Rice Krispies does not affect your ability to enjoy Corn Flakes. But the public policy problem is precisely that we have to chose one or the other! In that case, you might well argue that the best policy is the cereal that is liked by the most people, or the one that is most intensely liked, if we could measure that.

[82] As far as I know Myerson (1995) was the first to make this very basic and important point.

20 Multiple jurisdictions: To each her own?

Of course, it *is* possible to divide up political constituencies, so that they can have different policies or representatives for all who live in the constituency. This is part of the rationale for local control of things like school policies, zoning regulations, police departments, state income taxes and other tax policies. The observation is famously associated with Charles Tiebout's (1956) point that if people can sort themselves into political jurisdictions, they can select the public policy mixes of tax rates and public services that they like best.

For example, a political movement in eastern Oregon seeks to secede from Oregon and join into a larger state of Idaho. These are conservative voters who dislike policies chosen by the more liberal majorities who live mainly near the coast. They expect that they would be happier with the policies chosen by a Greater Idaho.[83] It seems entirely possible that such a jurisdictional change could at the same time improve the wellbeing of the average voter in the more liberal part of the state, by allowing policy to move to the left (because the state legislature would become more homogeneously liberal).

At the international level, Alesina and Spolaore (2005) develop related arguments concerning "the optimal size of nations." These are worth sketching as a way of better understanding the normative presumptions and implications of Downsian models of policy choice.

Consider a country, or a state within a federation, represented as the unit interval $[0, 1]$, in which there is contestation over cultural or regulatory policies – for instance, what curriculum should be taught in schools. Voters have ideal points distributed uniformly on the space, as in the Hotelling beach example. This jurisdiction chooses cultural policy by a majority rule election between two candidates, implementing the median policy $x_m = 1/2$. Assume that individuals' preferences over cultural policies can be represented by the component $-|x_i - x_m|$, which implies in turn that the average voter's welfare on the cultural dimension in the jurisdiction is $-1/4$.[84]

The voters must also pay taxes to fund a public good. To keep things extremely simple, let's suppose that all voters have the same pre-tax income, $y > 0$, and there is a single public good that must be funded and costs $G > 0$. Perhaps this is a fixed cost of setting up public schools or government offices, or it could be a cost of national defense, where G is the minimum needed to prevent invasion or coercion by a neighboring country.

With a linear (i.e., proportional) tax rate $t \in [0, 1]$, individuals have after-tax income $y(1 - t)$ and need to raise a total of G, which means that they need a tax rate t^* such that $t^*y = G$. We will represent i's preferences (conditional on the public good being funded) as

$$u_i(x, t) = y(1 - t) - a|x - x_i|$$

[83] Kirk Johnson, "Their Own Private Idaho: Five Oregon Counties Back a Plan to Secede," *New York Times* May 21, 2021. "Most of us in rural Oregon realize that whatever the Portland/Willamette Valley area wants, they get," Mr. McCarter said. "Do we have the freedom to vote who we want to govern us? That's the question." Scott Wilson, "Outflanked by liberals, Oregon conservatives aim to become part of Idaho," *Washington Post* September 15, 2023, documents the growth of this movement, and has suggestions that some supporters are concerned about the possibility of armed groups forming to seek secession with violent means.

[84] The average voter to the right of 1/2 is located at 3/4, and the average voter to the left of 1/2 is located at 1/4. Both get a payoff of $-1/4$ from $x_m = 1/2$.

where x is the cultural or regulatory policy; t is the tax rate, and $a > 0$ is a parameter that scales the importance of cultural policy relative to disposable income.

So, in the "unified" country or district with political competition producing $x = x_m$, i's payoff is

$$u_i(x_m = 1/2, t^* = G/y) = y - G - a|1/2 - x_i|.$$

Now let's consider an alternative arrangement where we have *two* jurisdictions, dividing this one in half at $x = 1/2$. Consider the left jurisdiction, $[0, 1/2]$. The new median voter is located at $1/4$, which means that the average voter's most-preferred cultural or regulatory policy is only $1/8$ distant from the policy implemented under majority rule. This makes for a welfare improvement of $a/8$ relative to politics in the larger jurisdiction. The reason is simple: Cultural or regulatory preferences are *more homogeneous* in the smaller jurisdiction, so that policy is on average closer to what people in the jurisdiction want.

But what about funding the public good? Notice that now the *tax base* has shrunk by one half – the same tax rate t now produces revenue of only $ty/2$ because there are only half as many producers in the new, smaller jurisdiction.

This is not a problem if government services are *constant returns to scale*, so that, say, halving the size of the population exactly halves the amount of funding needing to provide the same quality of per capita government services. But in many areas, there are *economies of scale* in public goods provision. For example, any time there are fixed costs for setting up and running government offices (or things like schools), there is a savings if you can spread the fixed costs over a larger number of tax payers.[85] Alternatively, consider national defense. To preserve its independence, a smaller country might need to spend almost as much as before the break up, if it needs to match the spending of an aggressive, nearby rival state.

To starkly illustrate the consequences of economies of scale in government, let's suppose that the new $[0, 1/2]$ jurisdiction still has to raise the same total for public goods, so that in the new, smaller jurisdiction, $t^* = 2G/y$. Then person i's payoff in the new set up is

$$u_i(x_m = 1/4, t^* = 2G/y) = y - 2G - a|1/4 - x_i|.$$

We can now see the tradeoff clearly. The average voter gains in the smaller jurisdiction in terms of left-right policy closer to her ideal point, but loses in terms of the per capita cost of government services or public goods. Dividing up the larger jurisdiction increases the welfare of the average voter only if $a/8 > G$.

Due to Alesina and Spolaore (2005), this type of model can be used to derive a rich and interesting set of implications. For example, lower levels of international security threat increase incentives for secession and the break up of federations (like Yugoslavia), by in effect reducing the need for a large tax base to fund high G defense spending. And the voters in the unified jurisdiction who most want secession and are most willing to accept the tax tradeoffs are those whose cultural or regulatory preferences are farthest from the median.

[85]For instance, while certainly you want more and bigger DMV's with a larger population, you can probably manage with DMV employment as a concave function of population size. Or consider that every household doing its own homeschooling is far more labor intensive that having schools with 20-30 students per teacher.

In domestic politics, almost all countries provide different government services at different jurisdictional levels, and important political contestation occurs over the allocation of responsibilities across levels and the drawing of internal political boundaries. We often see, for example, that richer communities want to – and often have – "seceded" from school districts so that they do not have to pay taxes to support public education for poorer residents. It is a worthwhile exercise to redo the above analysis without a separate cultural dimension of preferences, but instead just redistributive politics over a tax rate t that funds public goods to the tune of $t\bar{y}_\alpha\alpha$, where $\bar{y}_\alpha$ is mean income in a jurisdiction of size $\alpha \in (0, 1]$. A richer community within a larger jurisdiction prefers a lower tax rate (under typical assumptions), but may face a similar economies-of-scale in government services tradeoff.

Which policies should be set locally and which in higher-level jurisdictions? This question has been studied under the rubric of "fiscal federalism," mainly by economists (for a review see Oates, 1999). The basic principle is that policies that have either positive or negative externalities across units may be better decided at the national level. For example, at the core of the U.S. Constitution is Article I, Section 10, by which the federal government prohibits states from imposing any tariffs on trade with other countries or other states in the union, and also prohibits independent arming and pacts with foreign powers. The US Constitution is essentially, though certainly not only, a trade agreement.

...

21 Structures of policy preference

21.1 Is the median voter's favorite policy a good or a bad outcome?

As already noted, any policy *other than* the policy most preferred by the median voter makes a majority of voters worse off than they could be, as they understand their own preferences. If we think we should try to take into account intensity of preferences, whether on a utilitarian or a social-peace-while-deliberating rationale, then there is a similar argument in favor of the policy preferred by the average rather than the median voter (see below).

These are not complicated arguments. It is interesting that from Hotelling on, the standard view has been that it is clearly a bad thing for democracy if candidates and parties offer the same policy position, whether at the median or the average. The moral intuition is usually that full convergence would be bad because then voters have "no choice." Famously for US-based political scientists, the 1950 American Political Science Association's Committee on Political Parties' report stressed the importance of parties being able to commit to implement programs that offer voters "meaningful alternatives."

As we have seen, it does not work to try to justify this intuition with an analogy to consumer choice (Myerson, 1995). If we are constrained to pick one type of breakfast cereal for society (or for whatever the specific political jurisdiction is), then there is a strong argument for picking the one that maximizes average satisfaction, not one that is notably different from what most people want.

A related "no choice"-intuition sees full convergence as bad because it supposedly means that voters "could hardly be in a position to influence public policy with their votes" (McCarty, 2019, 21). But it

is precisely the fact that voters have a choice that is driving candidates and policy to converge on the maximally popular outcome in the baseline Downsian model.

Perhaps implicit in the "no choice" reaction is a sense that for democracy to work well, politics must *engage* citizens, and full convergence would mean that they would lose interest. People must think they have a choice even when in fact no individual voter is decisive and imagining that one individually has influence via the vote is a cognitive illusion. We (elites? everyone?) need to pretend that individual votes matter, and to make the pretense work we also need candidates and parties to commit to implement policy positions that are distinctly suboptimal. In my view this is a backhanded, odd, and sometimes paternalistic justification.

There are more elaborate arguments in support of having parties and candidates who favor policies that are, ex ante, not the most popular. We will see one in section 25.7: It may not be known what the typical voter wants, so that votes given to divergent party positions happen to help reveal this and so get policy closer to the true median. This is a plausible view, even if it has a strong "second best" quality and it is also unlikely that, at least in the US case, House district median voter positions move around anywhere close to the disparity between Republican and Democratic candidate positions these days.

A second possible justification sees divergence as good because it facilitates societal learning about what policies are most effective on dimensions of common interest like (say) economic growth. Often, conservatives and liberals are arguing that their preferred policies will, if implemented, be shown by experience to have better results for all, at least to some degree. I do not know of models that develop this idea, but I can imagine that if there is genuine uncertainty about what would work best on, for instance, health policy, society might learn faster from experience with distinct Left and Right policies than with middling Moderate policy. It is not clear whether this would be so under broad conditions, and the argument can work only for policy choices that are widely seen as means to an end that liberals and conservatives agree on.[86]

Academic treatments that present the median voter logic almost never comment on normative implications, except occasionally to intimate that full convergence is a bad outcome, as in Hotelling's "excessive sameness" view. Normally the focus turns immediately to the positive question of what might explain lack of full convergence.

At the same time, however, political scientists and economists clearly view "too much" divergence as a very bad thing! This gets called *polarization* and has become the subject of major concern in the US, not only for social scientists. Regarding politicians, fairly clear evidence shows that since the late 1970s, Republican and Democratic members of Congress have increasingly diverged in their policy preferences on the left-right dimension.[87]

Why is polarization bad? Political scientists and media commentary stress

- More gridlock in Congress that sticks in place old, status quo policies when there are probably needed reforms that majorities would prefer.

[86] [relationship to two-armed bandit problem. Banks and Sundaram?]

[87] McCarty (2019) provides a fantastic overview and analysis of research on political polarization. US politics is not my home field in Political Science, and I rely often on McCarty's synthesis presentation in what follows.

- "Unstable majorities" (Fiorina, 2017) that oscillate between Right and Left control, to the greater unhappiness of moderates and the out-party because the political winners have more extreme policy preferences, relative to the center and the other side.
- A toxic media climate and increasing mobilization of anti-system extremists, especially on the Right, which threatens democracy and social peace.

In each case, the outcome is seen as bad precisely because it involves getting or risking policies that diverge significantly from the median or average voter's preferences. So I take it that there is in fact strong agreement with the proposition that it is normatively problematic if a democratic system tends to produce policies that deviate markedly from what most voters think they want.[88]

We see the same presumption in the large literature on democratic representation [some cites]. Here, the usual normative standard is whether a member of Congress (say) votes in a manner more or less consistent with the preferences of the median or average voter in her district. It is assumed to be a bad thing – poor representation – if a representative's votes in Congress do not reflect how typical voters would vote on the bills (or perhaps, how they would vote if well informed).

So in what follows I am going to continue to use "distance from median or average citizen's preferred policy outcomes" as a serviceable normative standard for evaluating democratic performance.

21.2 Median versus average

But *which one*, median or average? And does it matter?

It matters if and when two conditions hold.

1. The median and average citizens' most preferred policies are different (obviously). If ideal points are distributed in a fairly symmetric way, the median and average are approximately the same. They differ only when policy preferences are skewed. Substantively this refers to situations where there is a minority that ideally would like a policy that is quite different from what most people would like best, and only "in one direction" from most people.

   To illustrate, consider Figure X, which reproduces Figure Y. This is the distribution of Left-Right ideal points estimated for US House members in 2016. The median House member was a fairly conservative Republican, while the mean member was a conservative Democrat. ... ?

   It is worth stressing that the normative importance of median-versus-average does *not* require a spatial framework. Consider a jurisdiction whose members are voting on whether to support a bond issue to pay for a bridge, or an increase in public school funding. Let $u_i \in \mathbb{R}$ be voter i's change in wellbeing if the measure passes. The measure passes if $u_m > 0$, where u_m is the

[88] In contrast to the literature on U.S. electoral politics, scholars of U.S. legislative politics at least implicitly associate bad performance of Congress with the passing of bills that are farther away from the median (e.g. Krehbiel, 1992; Weingast and Marshall, 1988; Cox and McCubbins, 2007). In a good example of the widespread rejection of explicit consideration of performance, Krehbiel (2010) is emphatic that he takes no position on whether the Congressional outcomes that this theory predicts are good or bad, even though in general his work seems to defend Congress against charges that it is not adequately majoritarian.

median value for the bill among the voters. But by the naive utilitarian standard, the measure should pass if and only if the average u_i, call it $\bar{u}$, is greater than zero. It can easily happen that the median voter doesn't like the measure ($u_m < 0$), but average wellbeing would increase if it passed ($\bar{u} > 0$). This kind of "tyranny of the majority" tends to occur when there is an intense minority who would greatly benefit from the bill while the majority is not much affected but has u_m slightly less than zero.

2. Preferences are such that people find policy that moves away from their ideal point increasingly more distasteful the farther away it gets. In other words, preferences are concave in the spatial dimension (or dimensions) in question. Quadratic preferences are an example.

   Suppose we move policy by one "unit" or step away the median voter's ideal point and towards the left. With linear preferences, the welfare loss of every person whose ideal point is between the median and the right extreme is the same as the gain of each of those on the left, regardless of how extreme they are. And since a majority loses, the median is the socially best outcome with linear preferences, by the naive utilitarian standard.

   By contrast, with quadratic or other concave preferences, if there are enough people with "hard left" policy preferences, then the net gain from getting closer to their preference can exceed the loss to a moderate majority. The "very left" were made *really* unhappy by the median voter's preferred policy, and there is not that much loss to the median and others nearby from a small move to the left. With quadratic preferences, the best policy by the utilitarian standard is specifically the average ideal point.

To see this formally, consider again the baseline Downsian model with voter ideal points distributed according to the cdf F. For an issue dimension on which voters' preferences can be represented by quadratic utility $u_i(x) = -(x - x_i)^2$, we can define the average societal welfare *loss* associated with implementing a policy $x \in \mathbb{R}$ by

$$\int (x - s)^2 f(s) ds.$$

For issues on which voters' preferences are better represented by linear utility, average societal loss associated with policy x is

$$\int |x - s| f(s) ds.$$

If you have taken basic statistics, you may recall that the x that minimizes the quadratic loss (or "mean squared error") is the mean $\bar{x} = \int s f(s) ds$, and the x that minimizes the average absolute difference is the median, the x_m such that $F(x_m) = 1/2$.

Because both mean and median are measures of central tendencies of a distribution, in public opinion on specific policies or on general left-right dimensions (social, economic, etc.), neither can be extreme. In both cases, then, *good democratic performance requires moderate policies*, as argued above. It is a question of how exactly to moderate.

Are policy preferences typically better represented as concave or linear?[89] This could vary issue to issue

[89] Or something else, such as convex or "Gaussian" preferences. For these, voters don't care about extremism beyond a certain distance from their ideal points. I return to convex and Gaussian preferences in section X below.

and probably also varies across people. I don't know and I am not sure if we have much systematic evidence.

I would think that regardless of their liberal-conservative preferences, most people would see alternation or randomization between their own ideal point and an equally conservative or liberal position on the other side as worse than just having the moderate policy in between.[90] Or in other words, if Left-wing Democrats see Centrist policies as bad, they are likely to see Right-wing policies as more than twice as bad as Centrist policies. And likewise for the Right-wing Republican, in the other direction. To the extent that this is so on general Left-Right or specific issues, the difference between what the median and average voter likes best can be both politically and normatively consequential.

Up to this point, we have been open to multiple interpretations of the single dimension of the baseline Downsian model. We have allowed that it could be thought of as (a) a general left-right, liberal-conservative, or strong-Democrat-to-strong-Republican orientation; (b) a specific policy issue, such as abortion, gun control, or average tax rate; or (c) something else, like a range of social types mappable to a left-right dimension of political preferences.

On the question of whether preferences are skewed so that mean and median are quite different, which interpretation matters. For specific policies, it is entirely common for a minority to care *a lot*, while there is a large majority with weak, vague, or uninformed preferences one way or the other.

By contrast, for individuals' general left-right orientations, on most measures there is not much difference between mean and median. With public opinion data in the US, when political scientists ask directly about left-right preferences (liberal-conservative "ideology") or about strength of party identification, we tend to get roughly symmetric bell-shaped distributions. When they employ various methods that use responses on specific policy questions to estimate an underlying or "latent" liberal-conservative disposition, they also get unimodal, roughly symmetric distributions.

This could mean that there are a lot of people who are on average rather moderate. Or it could mean that there are a lot of people who aren't paying close attention, so that their quasi-random answers are averaging to "moderation." Or it could mean that many people have distinct policy preferences across multiple issues that are just not very correlated with each other in a Left-Right or Democratic-Republican sense. Most likely there are people of all three types (Fowler et al., 2021). Regardless, there is not much indication of significant skew that would imply a significant gap between mean and median in terms of overall Left-Right policy preference, at least at the national level.

21.3 What's best when people differ a lot in what they care about?

The baseline Downsian model posits a single issue dimension when in fact governments choose and implement policies on many issues. What happens if we modify the model by introducing $k > 1$ distinct issue dimensions, and voter preferences represented by, say, additive quadratic or linear loss in the policy space $\mathbb{R}^k$?

[90] There is another consideration in regard to alternation, which is that there will almost always be direct costs of large policy swings and changes.

Regarding what is the best policy outcome given voter preferences, nothing changes! With quadratic loss, the naive utilitarian best outcome is now the average most preferred policy on each dimension, and with linear preferences it is the median on each dimension.[91]

But multiple issue dimensions do introduce a new consideration of great political consequence. As noted at the end of the last section, people can differ a lot in how much they care about different issues. This makes mean-versus-median conflicts common and significant at the level of specific policy issues. A large amount of what we call politics concerns conflict between "high demander" minorities, or "special interests," who want to move a specific policy towards what they really want, and majorities who have weak but opposed preferences on the specific issue.

For instance, US sugar beet producers want a prohibitively high tariff or other trade restriction on their import-competing product. If they understood the consequences for sugar prices and downstream economic effects, median voters might prefer zero or very low restrictions. The welfare-optimal policy might then be zero tariffs and some transfer to the sugar beet farmers, which might be similar in effect to a policy that implements the average most-preferred tariff (but more efficient).

Or, the abortion and gun policy issue areas are characterized by substantial minorities with intensely held preferences that are extreme relative to what the median voter prefers. These minorities exert costly effort mobilizing and lobbying politicians to move policy away from median-preferred policies, and they probably condition their votes on candidates' stances on these issues more than other voters do. The effect is to move policy away from the median voter and towards – or beyond – the average most-preferred policies.[92]

In fact we have already seen this kind of preference structure, twice. First, policy concerning the distribution of divisible material benefits. Second, the idea of dividing up jurisdictions to allow local

[91] In work on what they call "the delegate paradox," Broockman (2016) and Ahler and Broockman (2018) develop the following kind of example to argue that "polarized politicians can best represent constituencies comprised of citizens with idiosyncratic [i.e., "ideologically" unconstrained issue] preferences." Adam Schiff, Representative for a liberal Los Angeles House district (29th), votes for the liberal position on bills on five different issues (Fair Pay, SCHIP, Stimulus package, Clean Energy, ACA). This makes him appear to be "ideologically extreme" relative to voters in his district, since very few of his constituents express a preference for the liberal position on all five of these bills/issues. Scaling methods that estimate an underlying left-right dimension from these bills would locate Schiff on the far left, while the same methods would locate the large majority of his constituents as more moderate.

This example nicely identifies a problem with our scaling methods, but it should be stressed that it does not mean that extremists can be better representatives when there are multiple issues. (Ahler and Broockman do not say this, but it may be easy to miss.) In the spatial framework presented above, Schiff's votes make him an unambiguous moderate in his district. If we imagine that there really are just two possible policy positions on each of these five issues (left or right), then Schiff is adopting the position of the median voter on all issues – you can't be more moderate than this (if Schiff voted the conservative position on any one bill, there would be a majority of his constituents who disagreed). If by contrast we imagine each issue dimension as more continuous, then Schiff's votes still imply that he is taking the position closer to his constituents' median on each issue. If bills in the House presented Schiff with the opportunity to vote on versions that are too far left for his median constituent, then the scaling method would correctly locate Schiff's positions in the center for his district (if he was being a good representative in this sense). The problem here is that the House agenda does not provide enough variation in bills to properly identify cutpoints for representatives and constituents with positions away from the center. This problem tends to exacerbate the perception of polarization by making ideal points "clump" more within parties, even as it underestimates heterogeneity.

[92] Section X considers a model in which candidates compete by offering "transfers" to specific groups, which can also be interpreted as favorable policies on particular issues that the groups care about. A straightforward result from that model is that relatively "single-issue voters" get catered to more than others, because their votes are very responsive to candidate positions.

policy choices to better match local policy preferences in cultural, regulatory, and other matters.

- Recall the problem of choosing how to distribute a fixed amount of income or other material benefits across n people, all of whom like more rather than less and put zero (or very little) weight on what others get. Say the total amount to be divided up is $B > 0$. It is just as if we have n distinct policy dimensions representing how much each individual gets, and ideal points such that i likes getting $x_i^* = B$ the most and doesn't care what others get unless it affects her amount.

  We saw that with concave preferences, the (naive utilitarian) optimal policy equalizes welfare gains from additional stuff across people, which implies an equal distribution when $u_i(x_i)$ is taken to be the same for all i. In this kind of strategic situation, giving more stuff to one person (or household, or congressional district, ...) has a "negative externality" for all the others. Among other things this makes it sensible to decide on the distribution centrally, rather than allowing the units to draw from a central pot without consideration for other districts or units.[93]

  One major function of elected legislatures is to decide how to allocate spending on public goods across legislative districts – roads, bridges, and defense contracts, for instance. Such *distributive politics* problems are characterized by a tension between majority-preferred extremism – zeroing out a minority of districts to allow more stuff for members of a majority coalition, and policy moderation in the form of a "universalist" coalition that distributes goods more equally (Weingast, 1979, 1994).

- For a range of other specific policy matters, it is possible in principle for one jurisdiction in a country to have a different policy from others, without there being substantial negative externalities. If a large majority of people in District A don't want liquor sold on Sundays, but almost everyone in District B is completely fine with this, then there is a plausible argument that it is better to allow each district to choose its own policy on this matter rather than to try select a single policy for both.

  Likewise at the global level. Having an effective, well-functioning world government might be welfare-improving on a host of policy matters where there are major negative externalities due to national policies. A good world government might make possible lower carbon emissions, less war, less risk of nuclear war, and greater economic prosperity. But it would be extremely difficult to get any agreement on an effective world government unless existing countries felt assured that they could not be forced to, for instance, impose a state religion chosen by the international legislature. An international or regional federation either needs to allow considerable "district"-level choice in a range of policy areas, or be prepared to deal with significant threats to social peace (in the sense discussed in section X above).

  Returning to blue laws, what if people in District A, who don't want liquor sold on Sunday, believe strongly that this should also be the case in District B? Or what if majorities in some US states want more restrictive laws on abortion or gun control than majorities in other states? If people in one jurisdiction have strong views on what the policies should be on particular issues in *other* jurisdictions, then state- or district-level policies by definition can have negative

[93] A common problem in federations is that the lower-level units (states, provinces) have some discretion to spend and the central government can't fully commit not to bail them out if they overspend (Rodden, 2006). [the problem of soft budget constraints ...]

externalities outside the district.[94]

What constitutes "negative externalities" that would favor policy choice at the national level is thus itself a function of the structure of preferences on the specific issue in question. The multiple-jurisdictions solution to matching policy to preferences on multiple issue dimensions depends on either pragmatic acceptance of, or moral agreement on, tolerance of what "those other people" prefer.[95]

For later use let's formalize additive, quadratic-loss policy preferences with multiple issues. There are $k > 1$ issue dimensions with policies on them represented by real numbers, thus a policy space in $\mathbb{R}^k$. Let voter i's ideal point be a point in this space, $x_i \in \mathbb{R}^k$, with x_{ij} standing for i's ideal point on issue j. Consider a vector of policies $x = (x^1, x^2, \ldots, x^j, \ldots, x^k)$, where we use superscripts to indicate the policy choice on issue j as opposed to a voter's ideal point (subscripts). Voter i's preferences would then be represented by

$$u_i(x) = -\sum_{j=1}^{k} a_{ij}(x^j - x_{ij})^2, \tag{5}$$

where preference parameters $a_{ij} \geq 0$ scale how much i cares about issue j relative to the other issues.

To weight each person's wellbeing from policy equally we might fix $\sum_j a_{ij} = 1$ for all i.

Even with such an assumption, however, the informational demands to identify optimal policies in this set up are truly forbidding. This is so for several reasons.

1. Very few people have well-developed and well-considered preferences over policy dimensions that do not directly affect their daily lives. More discussion in the next section.

2. On specific policy dimensions, *some* people, or businesses, or interest groups, often have strong preferences (large a_{ij}) and, typically, ideal points that are extreme relative to a large majority who are not so directly concerned or affected. This preference structure makes for a mean-median conflict when the "high demanders" are mainly on one side of the issue. How then to determine optimal policy, given that no one can simply observe how much the high demanders care relative to other issues and relative to the median citizen's preferences on the issue? In practice the answer is lobbying and other forms of costly mobilization by high demanders. Politicians then trade off benefits from helping the high demanders against costs of going against majority preferences or interests.[96]

   For other policies, there may be intense "single-issue voters" on both sides, which tends to favor more moderate policies under a range of assumptions about lobbying and political process.

[94] In the case of gun laws, these externalities can include more gun violence in a jurisdiction that borders a neighbor with loose gun laws (Dube, Dube and García-Ponce, 2013).

[95] The political philosophy of classical liberalism is said to have been born out of pragmatic acceptance of Protestant-Catholic religious tolerance, both within and across European states. After the US Civil War, the Northern states' representatives backed off Reconstruction in the face of the threat of terrorism and guerrilla war by white Southerners, a pragmatic acquiescence to immoral policies (racial authoritarianism and violent repression) favored by powerful majorities in some states.

[96] Models of this type of strategic situation are most developed in the area of lobbying for trade restrictions (Grossman and Helpman, 1992). More generally, on "special interest politics," see Grossman and Helpman (2001).

3. By assuming an additive structure, the formulation of policy preferences in (5) is already vastly too simple. What about tradeoffs and interactions across policy domains? Choosing a policy on one issue often, if not always, has implications for other issues. These interactions are complicated, requiring investigation and expertise to assess or to make informed guesses about.

   The complexity of the substantive issues and tradeoffs between policies is a major reason for having representative rather than "direct" democracy. It is also a major reason why it makes far more sense to evaluate democratic performance based on policy outcomes rather than measures of congruence between survey responses and measures of politicians' position-taking.[97]

Voting systems can be devised that in principle could do a better job than simple majority rule by revealing relative intensities of public preference across different specific policy questions. Lalley and Weyl (2018) propose *quadratic voting*, in which voters "buy" votes on different Yes/No policy questions by allocating tokens, and the cost of votes increases quadratically in tokens (e.g., one vote on an issue costs 1 token, two votes on that issue costs 4). Such a system allows voters to express relative intensities of preferences across issues. The informational demands of such a system for voters in mass elections are very high (see points 1-3 above). We briefly return to analysis of mean-median conflicts as a mechanism-design problem in section 32 below.

21.4 Do voters even have policy preferences? Should they?

If government were as simple a matter as metaphorically choosing a point on a liberal-conservative line, it would probably look very different. Instead, governments formulate and select policies on many dimensions. Even in one issue area, any one policy normally involves tens of thousands of details and decisions when considered from formulation of a bill through to implementation.

As we did above, in principle we can abstractly represent multiple dimensions and specific policy details by going from a policy space of $X = \mathbb{R}$ to $X = \mathbb{R}^k$ where there are $k > 1$ dimensions. But it is completely absurd to imagine that citizens have awareness of the existence of, let alone well-formed preferences over, all these policy dimensions.

A fundamental rationale for *representative* democracy is that citizens in effect hire politicians and bureaucrats to specialize in researching, bargaining, and doing the detailed work of formulating and implementing policies. It is bizarre to imagine that people would have well-thought out preferences about optimal policy on a host of matters that do not directly affect their daily lives.

What to make, then, of the democratic theory embedded in the baseline Downsian model, that implies that we should judge democratic performance by the congruence between citizens' policy preferences and the policies that are chosen by politicians and implemented by bureaucrats?

If we replace "citizens' policy preferences" with "citizens' policy preferences *if they were well-informed about the specific policy questions at issue*," then this performance standard remains compelling. If you are choosing a mechanic to work on your car, you want someone who will make the decisions that you

[97] Regrettably, for the most part political science long ago ceded the study and evaluation of US government performance on policy outcomes to economists and sociologists, in favor of studying survey responses, vote choice, and, for a time, congressional organization.

would make if you were well informed about car engines. And good performance concerns whether the mechanic in fact makes the decisions that are close to what you would want.

Lacking policy expertise and all kinds of relevant knowledge, what are voters to do? They have two main avenues of recourse.

1. Punish incumbent politicians based on observations of how things are going, including if they observe policies or possible effects of policies that they don't like. This is known as *retrospective voting* and is the subject of models considered later in the book. For retrospective voting, the question of democratic performance becomes whether and how well voters can monitor and condition their votes on politicians' choices and government performance. (And what can be done to improve this.)

2. Prospectively, vote based on candidate qualifications and experience, but as important, on whether the candidate seems like a comprehensible social type who has interests and inclinations similar to one's own. Is he or she a member of my group? Is he or she "people like me"? Note that voting on the basis of impressions of social type and group membership makes great sense when we consider that it is extremely difficult for voters to evaluate what individual politicians do when they are in office. This is a principle barrier to effective retrospective voting, and it pushes voters in the direction of trying to identify social types they believe they can trust to have similar policy preferences (Fearon, 1999).

In other words, *social types proxy for policy preferences*.

In the US, the main parties bundle sets of social types, on the one hand, and sets of policy preferences on the other. The latter are strongly influenced by interest groups that really do have specific policy preferences. For issues that individuals don't have a particular reason to learn or know about, they guess at their policy preferences based in part on imperfect knowledge of party policy bundles. Which party they identify with (and how strongly) is based on knowledge of which social types bundle with which party.

Many expressions of policy preference on specific issues may thus result from a conscious or unconscious logic along the lines of: People like me are (say) Republicans, and I think that Republicans like policy X, so I probably prefer policy X.

We should also expect a fair bit of: People like me are Republicans, I think policy X sounds good (based on wording of a survey question), so probably this is also what the Republican party likes. This is a reasonable inference for a low-information respondent.

What follows for evaluating democratic performance? This question is underexplored, I believe.

One obvious implication is that voters' self-reports of their policy preferences on specific issues come with error attached, relative to what the person would prefer if more fully informed. Larger average error should be expected for policy matters of low salience, and in areas where party policy caters to interest groups whose preferences are not well aligned with what party supporters would want if they knew more about what was going on.

It is probably wrong, however, to completely disregard average self-reported policy preferences on specific issues. For one thing, the inference from beliefs about party X's policy positions to one's own likely policy preferences is not unreasonable. If you are the type of person who tends to support party X, then politicians from party X have an incentive to pursue policies that would have results that social types like you would prefer! This is clearly not a perfect mechanism, as discussed in the sections on retrospective voting and the electoral control of politicians. But evidence that people infer their policy preferences to some degree from beliefs about party positions does not imply that these guesses are wrong on average.[98]

Returning to the one-dimensional Downsian model of a general Left-Right dimension, we can think of this as a spectrum ranging from very "liberal" (or Democratic) to very "conservative" (or Republican) social types, on which people don't seem to have too much trouble placing themselves. People get vastly more day-to-day information about social maps and social types than they do about the details of, say, trade policy.

21.5 Social dominance preferences and descriptive representation

Given the strategic situation, it can make sense for voters to use social type as a cue to try to identify politicians who might act on their behalf in policymaking. But what if the social type of their representatives is not just an indicator. What if, in addition or instead, it is a source of satisfaction by itself? What if, independent of government performance, it simply makes some voters happier to have a same-race, same-ethnic group, same-religion, or same-gender political representative?

Political power implies hierarchy. Those who hold it give the commands. So it is not surprising that people derive vicarious pleasure from seeing "people like me" at the top.

It is debated, however, whether this kind of preference *should* get weight in democratic governance. On one view it is just wrong to take pleasure or pride in "ascriptive" qualities over which we have no control, so that a pure, non-instrumental preference for own-ascriptive-group representation doesn't merit and shouldn't get public respect. (The preference is non-instrumental if own-group representation is valued for its own sake. It is instrumental to the extent that social type is used as a means to identify a representative who is more likely to choose policies that are what you would like, or that would be in your well-considered interest.)

Others might distinguish between a pure preference for own-group representation based on

(a) a belief in the intrinsic superiority of one's own group, which is imagined to justify social and political hierarchy, and

(b) the belief that pleasure from own-group representation is natural, or at least ok. But if one wants it for one's own group one should think it fair that it be shared with members of other groups.

[98] A large literature and long tradition in the field of Political Behavior assumes that democracy is not working if people often derive their policy preferences as expressed in surveys from beliefs about party positions; see, for example, Achen and Bartels (2017); Lenz (2013). The despairing view about the possibility of democracy in this tradition (people aren't up to it) is premised on an implicit acceptance of the Downsian image of elections translating well-formed policy preferences to policy through candidate position-taking in elections.

Examples of (a) include white supremacism or any ethnic nationalism premised on an idea of intrinsic racial or ethnic superiority. Because they reject political equality such beliefs are poisonous for democracy when they are held by some citizens about other citizens. It would seem contradictory, or certainly not conducive to social peace, to value satisfying them as a policy objective.[99]

The idea of *descriptive representation* – which refers to the degree of similarity between the social type of the representative and his or her constituents – can be an example of (b), although arguments for descriptive representation are usually instrumentalist. That is, they offer reasons for why having a representative who "looks like me" serves ends like democratic legitimacy, trust in government, policy implementation, and better representation of interests (as argued above; cites). But one could allow that there is just some intrinsic value in seeing "people like me" in higher office if it is not premised on an equality-denying ideology. If so, the question becomes how to optimally distribute or allocate this "indivisible" good.

Let's consider the extreme case of a society or a legislative district with two ethnic groups whose population shares are p_1 and p_2, with $p_1 > p_2$ (and $p_1 + p_2 = 1$). As a thought experiment we will imagine that members of each group care *only* about the social identity of a single representation like the president of the country or their legislative representative.

Fix wellbeing for our naive utilitarian social welfare function so that each person derives a benefit of 1 if the representative is from their group, and has a value of zero if not. Suppose that we can design a political system that has the effect of selecting a person from group i to be president (say) with probability r_i. In the simplest case this can just be a lottery r with probabilities $r = (r_1, r_2)$. What is the socially optimal lottery r, given that we take each person's loss to be the same for not having an own-group leader?

This is an interesting case where what seems fair is different from what is best from a utilitarian perspective. Intuitively, the fair thing would be to set $r_i = p_i$ so that each group's probability of being descriptively represented is equal to its population share. But then average expected utility is $p_1(p_1 * 1 + p_2 * 0) + p_2(p_1 * 0 + p_2 * 1) = p_1^2 + p_2^2$. If group 1 is a strict majority, then this is less than p_1, which is the average value for just making a group 1 person president for sure. Since there are more members of group 1 than group 2, the naive utilitarian best policy is to maximize welfare by satisfying the majority preference.[100] If you have to chose which of two people to give an indivisible object to, should you not give it to the person who values it more? The logic here is similar, counting satisfaction with the policy equally across individuals.

Indeed, in other examples with an "indivisible" or binary policy choice at issue, satisfying the majority's preference can seem like the best option. If for some reason a community is constrained to have just one cereal made available, Rice Krispies or Corn Flakes, then choosing the majority-preferred cereal is the more moderate policy. And likewise for any more real-world policy choice that, for some reason, has to be one thing or another.

What to do? The difficulty is that with binary options, there can be a stark conflict between fairness, which requires that everyone get a(n expected) share of the benefits, and the utilitarian principle of

[99] Somewhat reassuringly, Ansolabehere and Fraga (2016) examine U.S. survey data and find fairly slightly indication of pure preference for co-ethnic representation after taking party identification into account.

[100] With this binary outcome, it does not matter if we represent preferences as linear or quadratic.

maximizing aggregate wellbeing. With just two possible choices, there is no hope of a plausible concavity assumption softening the conflict, as occurred with concave preferences over income (section 17).

21.6 Convex preferences and social peace considerations

Let's consider another example of polarized preferences across social groups in an electorate. Imagine a congressional district that is 60% very conservative Republicans, and 40% very liberal Democrats. Allow that each voter's value for getting a representative with equivalent political and social views is 1, parallel to our above example of descriptive representation of ethnic groups.

If there are two candidates for office, one very conservative Republican and one very liberal Democrat, then the situation seems at first glance rather similar. The naive utilitarian best outcome is that the Republican candidate holds office with probability 1, which shuts out or fails to represent the 40% minority at all. Intuitively a more fair outcome would be if the Democratic candidate had a 40% chance of being elected, or perhaps won office in 40% of races over time.

But the situation is *not* the same, because it is easy to imagine that a political moderate could run for office, whereas idiosyncratic social conventions in the US and many other countries understand ethnic and racial identities as discrete things (an equilibrium in a coordination game that structures social maps and self-understandings!). Consider Figure X, which depicts possible policy preferences for an extreme liberal and an extreme conservative who live in this hypothetical district, defined over a continuous ideological policy dimension represented by $X = [0, 1]$.

Suppose first that all Republican voters have quadratic preferences on X represented by $u_R(x) = 1 - (1 - x)^2$, so that they all have the same "ideal point" at $x = 1$. All Democratic voters have preferences represented by $u_D(x) = 1 - x^2$, so their ideal point is at $x = 0$. As illustrated, these preferences are concave (curving down) in X. If the share of Republican voters in the electorate is p_R, then the naive utilitarian optimal policy is the x that maximizes $p_R u_R(x) + (1 - p_R)u_D(x) = p_R(2x - x^2) + (1 - p_R)(1 - x^2)$, which you can work out to be $x^* = p_R$. Once again, with quadratic preferences, the naive utilitarian best policy is the mean of the ideal points weighting each person equally.[101]

With 60% "severely conservative" Republicans, the optimal candidate for average wellbeing is thus slightly right of center ($x^* = .6$) but not nearly as extreme as a candidate who matches the median voter's views ($x = 1$). Concavity represents preferences such that the gain in wellbeing for the very liberal Democrats of moving a little towards $x = 0$ from $x = 1$ is considerably larger than the loss to the conservative Republicans, enough to partly compensate for the smaller share of Democrats in the district. With quadratic preferences on a sufficiently divisible set of policy options, the naive utilitarian best policy is that preferred by the *average* voter (who in this example is a hypothetical "expected" voter).

In principle, however, voter preferences over Left-Right policy positions need not be concave. Figure X also displays an example of convex preferences. The specific example shown is $u_R(x) = x^2$ and

[101] Exercise: Work out the utilitarian best policy for the following preferences: $u_R(x) = x^\alpha$ and $u_D(x) = (1 - x)^\alpha$, where $\alpha \in (0, 1)$ is a risk-aversion parameter (lower α, greater risk aversion). Compare your results to the quadratic case and explain.

$u_D(x) = (1 - x)^2$.

Exercise 30. *Show that the naive utilitarian best policy in this case of convex preferences is the ideal point of the more numerous group. Then prove that this holds for any strictly convex preferences, $u_R(x)$ and $u_D(x)$. (Hint: Think about the Figure.)*

Substantively, people with convex preferences prefer taking a 50-50 gamble on their ideal point and the other extreme, versus compromising on a political moderate located at $x = 1/2$. In other words, they don't see a big difference between moderate policies and the extreme that they dislike, relative to their most-preferred option.

Are convex preferences empirically common in political contexts? Are there a large number of people who would sincerely prefer that policy alternate between extremes versus a stable moderate policy, or a coin toss to decide who gets their extreme preference? I am doubtful myself, but let's continue to imagine a country or a district whose citizens are all divided in this way and who have these preferences. As we have seen, the naive utilitarian best policy is then the policy preferred by the majority, which seems unfair for the excluded minority.

Indeed, some members of a very unhappy 40% minority might be disinclined to live peacefully with the majority's most preferred policy (or policies), particularly if what is at issue is seen as an important moral disagreement and thus not simply a matter of satisfying different tastes. Demonstrations, protests, boycotts, terrorist attacks, or organized armed rebellion – by such means members of the dissatisfied minority may be able to threaten costs that would incline the majority to prefer a more moderate policy to trouble. These means may be unjust, but it could also be that the majority's preferred policy is unjust. In such circumstances a pragmatic compromise while the moral disagreements evolve could have a reasonable claim on being a good policy, or at any rate better than the effects of trying to impose either extreme (even if you think one side happens to be on the side of right).

Exercise 31. *For example, suppose that the minority has an option to use force to try to seize power, which it can do with probability w but at cost $c_L > 0$ and $c_R > 0$ in the fight that would occur. (a) For concave preferences, show that there is a set of compromises between the extremes that both sides prefer to the fight. (b) For convex preferences, show that there is a set of lotteries on the policies $x = 0$ and $x = 1$ that both prefer to the fight.*

In sum, with strictly concave preferences, moderation is recommended as good policy by both a utilitarian and a social peace rationale, even when popular preferences are highly polarized. With convex preferences, moderation in the form of alternation or lotteries on the extreme, polarized alternatives may be the best feasible policy under a social peace rationale.

21.7 But what if the typical voter is just wrong?

Among other things, in this section I have argued that a good standard for good performance of a political system is that it selects and implements policies that are close to what the typical citizen wants, or would want if well informed about the relevant facts and tradeoffs. You may object, "But on issues X, Y, Z, ..., I think that what the typical voter wants is just flat out bad and wrong. I do not agree that what the typical voter wants in those cases is 'good policy.'"

Well, yes. And you are not the only one! Lots of citizens are not typical (average, median).

You may well disagree with the median or typical voter's preferences, but that is a separate issue from the question of what policy a well-functioning democratic system should select. The following are two *different* questions, even though political scientists and others often seem to assume they are the same.

1. What policies do *I* think the government should implement? (For instance, Left, Center, or Right.)
2. What policies, if implemented by the government, would be most consistent with democracy performing well? (L, C, or R, say.)

If you answer L or R to the second question, consistent with your personal preference, then you should also be able to explain why or how you think it is maximally democratic to implement policies that are preferred by a minority when there is a majority-preferred alternative available.

There are ways to do this, such as

- "The media is controlled by R (L). What the median voter would want if they were not propagandized by R (L) is L (R)." An obstacle for this line is that people can choose which media to get their information from to a significant degree. Pursuing this line tends to lead in the direction of "the people must be forcibly reeducated by a right-thinking government."
- "If the median voter experienced L (R) policies, they would learn that they like them." An obstacle for this line is that (to this point) party control in the US government *has* varied over time. There have been occasions when L or R's preferred policies in some domain or domains have "taken," at least for some time, as with a number of the central New Deal initiatives. But there has also been a strong tendency of centrist voters to swing away from the party that gains more control of the presidency and Congress (see section 23).[102]

Would you argue that good performance means choosing – and imposing – policies that you think are morally or practically best, even when these are disliked and opposed by large majorities, or are opposed by passionate minorities against largely indifferent majorities? How would you reconcile your disregard of the democratic principle in the first case, when you probably are otherwise outraged at democratic violations that occur when minority preferences prevail over what you think is right when you are in the majority?[103]

In a healthy democratic system, those who disagree with centrist policies have means to try to persuade people to change their minds, and means to mobilize support to try to change policy. But they may not be successful. Unless everyone has the same views on all policies, which is difficult to imagine for any society, any political system has to select policies that some think are bad or wrong.

[102] In limiting cases this argument can appear as "true communism has never been tried"-type claims.

[103] Another way to put this: These two questions are different, although people tend to conflate them. (1) What policy do you personally think would be best? (on guns, or abortion, or taxes ...) (2) What is the policy that, if chosen the political process, would represent good democratic performance? (and given that people disagree and have diverse view)

22 Good and bad outcomes with multiple issue dimensions and multiple districts

22.1 Welfare loss = policy bias + preference heterogeneity

Consider again Figure 11(c). For average voter wellbeing, it won't matter much whether federal policy on some issue lands at the mean versus the median. They aren't that different. Nonetheless, because there is a broad range of policy preferences in the population, a lot of people are still not going to be happy with government policy, even if it is close to the mean or median.

By contrast, looking at Figure 11(d), if Republican control of Congress led to choice of a federal policy favored by the median Republican (blue dotted line), that policy is significantly away from the average in Congress and in the voting population. In this case, there is going to be substantial voter unhappiness related to both preference heterogeneity and what we will call "bias." Bias can be understood as a kind of *democratic failure*, in that a democratic system is producing policies that large majorities see as worse than other available policies.

Our formal framework gives us a way to decompose these two types of dissatisfaction with government policies. Consider a continuum of citizen ideal points distributed by density function f, with mean $\bar{x}$, median x_m, and variance σ^2. With quadratic preferences and policy at x, the average welfare loss in the population is

$$L(x) = \int (s - x)^2 f(s) ds = (x - \bar{x})^2 + \sigma^2$$

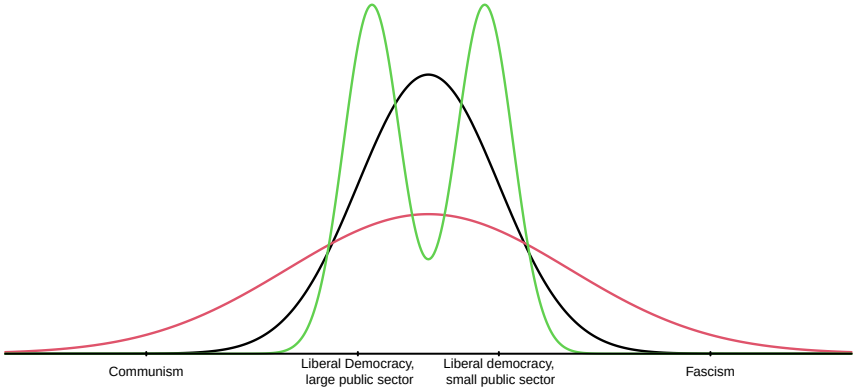
$$\text{average welfare loss} \;=\; \text{policy bias} \;+\; \text{preference heterogeneity}.$$

This is the bias/variance decomposition from basic statistics. The best policy with quadratic loss is what the typical citizen likes best, where "typical" in this case means average, not median. But even with the best policy, a highly heterogeneous population will still have a lot of unhappy people.

This decomposition does not depend on there being just a single Left-Right policy dimension. It works the same way if we have $k > 1$ policy dimensions represented by $\mathbb{R}^k$ and a density function $f(x)$ giving the distribution of voter ideal points $x_i \in \mathbb{R}^k$ in this space.

Empirically, Brunell (2010) found that US voters in more heterogenous districts are more unhappy with Congress. He argues that more homogeneous and thus less competitive districts (competitive between the parties) are actually better because greater preference homogeneity increases citizens' satisfaction with their representatives, on average. On similar lines, Stephanopoulos (2012) found that more "spatially diverse" districts with more heterogeneous populations have higher "roll off" rates, a measure of disengagement, and also a weaker connection between constituent preferences and representatives' voting in Congress. Although these works focus on how best to redistrict, the core point is the same as that made above by Alesina and Spolaore (2005) and, perhaps, advocates for Greater Idaho (section 20).

Figure 12: What is polarization?



22.2 What is polarization and what is ideological constraint?

The decomposition can also help us to investigate the concept of *political polarization*. The term has a great deal of intuitive and normative resonance, but it also turns out to have multiple plausible interpretations, or aspects.

1. Looking at the decomposition, we might associate polarization with the heterogeneity of political preferences in the population (σ^2). More *heterogeneous preferences*, greater "polarization." As an example, consider Figure 12, which represents three distributions of preferences over a Left-Right spectrum running from Communism to Fascism. It makes good sense to describe the society with preferences represented by the red line as more polarized than the society with preferences represented by the black line. It has a substantial number of people who favor Communism as well as many who favor Fascism. Think Weimar Germany, for example. In the black-line society hardly any one supports these very different regime types and there is something of a consensus on liberal democracy. The difference between the distributions is greater variance of ideal points in the former.

2. Consider again the "liberal consensus" society of the black line. Imagine that the distribution of preferences over the size and functions of the public sector remains the same, but that people become *more intense* about these preferences. That is, suppose the welfare loss perceived by a liberal (in the left-y sense) from conservative policies increases, and likewise the welfare loss

perceived by a small-government conservative from left policies increases. Then the aggregate societal welfare loss from any policy increases, as does the loss for any given amount of policy bias away from the mean. So it is again natural to say that "polarization" has increased, even though in this case no one's ideal point over policies has changed.

This idea, or type, of polarization can be captured with a simple modification to the expression for social welfare loss above. Let individual i's welfare loss from policy x be $a(x - x_i)^2$ where x_i is her ideal point and $a > 0$ scales intensity. Then societal loss becomes $a(x - \bar{x})^2 + a\sigma^2$.

Regarding public opinion, this kind of polarization will be completely missed by standard approaches to measurement, which target type 1 polarization by asking about a respondent's most preferred policy, or with questions like "generally speaking do you consider yourself a liberal, a conservative, or neither, ..." (or ditto for party identification). The distribution of peoples' preferred policies could stay constant, while they become more intense in their dislike of other policies.

Although the concept is not the same, type 2 polarization is plausibly related to the idea of "affective polarization," which refers to the intensity of individuals' dislike of members of the other party in US politics (Iyengar et al., 2019). Affective polarization might be represented by – or be correlated with – the degree of unhappiness people derive from policies that they associate more closely with political others.

3. Now consider the distribution of political preferences represented by the green line. Intuition almost surely says that this society is more politically polarized than the nice, normally distributed black-line society. But measured by the variance σ^2, preferences are actually *less* heterogeneous in this bimodal distribution! The variance is one-third smaller. Worse, imagine a society equally divided between people with ideal points heaped exactly at the indicated "small public sector" and "large public sector" positions. We would surely say that that society is more polarized than the black-line society, but the variance of political preferences is half as large.

   The problem is that heterogeneity does not uniquely capture the idea of polarization. "Polarization" contains also the idea of *poles*, or clusters that are internally similar but dissimilar to other clusters. Esteban and Ray (1994) proposed that polarization is greater (1) the greater the degree of internal homogeneity within groups; (2) the greater the degree of heterogeneity across groups; and (3) the smaller the number of significantly sized groups. Producing a single numerical summary measure of "polarization" so understood is difficult, and the details are not immediately relevant for present purposes.[104]

   For present purposes, observe that increasing polarization of the sort seen in the bimodal distribution of Figure 12 could either increase or decrease average welfare loss $L(x)$ for any given policy x, depending on whether it associated with increased or decreased variance. It is true that we might suspect that green would be prone to greater policy *bias* than black, if electoral and legislative institutions confer some power to choose policy on a majority winning coalition, separate from the median voter in the society as a whole (see X below for more on such legislative politics). And perhaps within-cluster or within-group similarity also facilitates collective action and mobilization, making for more resources wasted on conflict (Esteban and Ray, 2011). But

[104] See Esteban and Ray (1994); Duclos, Esteban and Ray (2004) and related literature. Studying party polarization in Congress, Poole and Rosenthal (2000, 81) agree with Esteban and Ray: "Polarization thus has two highly related, but distinct, aspects. For parties to be polarized, they must be far apart on policy issues, but be tightly distributed around the party mean." They measure and analyze the two aspects separately.

altogether it is not obvious why polarization in this sense would necessarily associate with lower average welfare, separate from any effect of preference heterogeneity.

4. The last three aspects of "polarization" were explained by reference to a single policy dimension, but each applies in exactly the same way for a distribution of policy preferences on $k > 1$ dimensions. With multiple dimensions, however, a *fourth* defensible interpretation of "polarization" arises. The idea is illustrated in the left panel of Figure 13, which depicts two issue dimensions, x_1 and x_2, and two sets of voter ideal points in the issue space. The dimensions could be, for example, Left-to-Right on size of government and functions of the public sector (x_1), and Left-to-Right on social issues (x_2, higher is more conservative). For a point of comparison, the right panel shows Larry Bartels' empirical estimates of voter positions on these two dimensions, based on his combining of survey responses to multiple issue questions for each dimension (Bartels, 2018).[105]

   Notice that the ideal points of individuals in the gray-dot society are almost uncorrelated across dimensions. Knowing a person's position on size of the public sector (redistribution, public goods) is not very informative about their position on social issues (gay marriage, abortion, gun control, drug policies, ...). By contrast, in the black-dot society, positions on the two issue dimensions are highly correlated, so that if you know that a person is, say, a liberal on redistribution/tax policy, you can be fairly sure that they are also socially liberal.

   In political science jargon, voters in the black-dot society exhibit a high degree of *ideological constraint* whereas ideological constraint is very low in the gray-dot society (Converse, 2006).

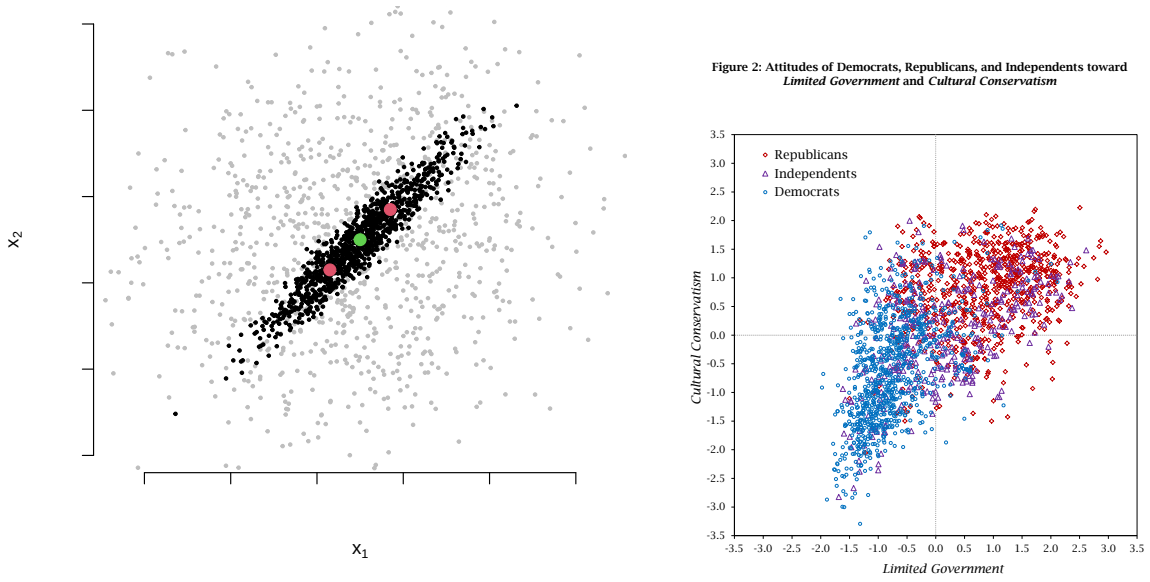
   It is important to flag that the idea of "ideological constraint" has nothing to do with the logical coherence of an ideological stance, but instead is purely about how well correlated voters' preferred policies are across different issues, when these are ordered by some analyst's notion of what is more "liberal" or "conservative." For example, there is no particular logical reason that favoring low taxes and little redistribution should imply that one would also favor socially conservative regulatory policies, as happens in the black-dot society. To the contrary, on a number of social issues the more logically coherent ideological position is probably libertarian, wherein a person prefers minimal state control and intervention in both the economy and people's private lives.

   Returning to "polarization" and heterogeneity, policy preferences in the gray-dot society are obviously much more heterogeneous than in the black-dot society, so any policy pair $x = (x_1, x_2)$ is going to imply lower average social welfare in that population. However the black-dot society also appears to be – or could be – significantly more polarized, because there is a much greater homogeneity of preferences within potential political coalitions. Imagine that two parties campaign on and try to implement the positions indicated by the red circles. Probably we would say that there is then greater polarization in the black-dot society than the gray-dot society, given that there is much stronger clustering of public preferences around the two party positions in the former.

   "Ideological constraint" makes for a kind of homogeneity of preferences, which as we have seen also increases average satisfaction with any given policy outcome (relative to lower constraint,

[105] By combining respondents' answers on several issues into single cultural and economic dimensions, Bartels' procedure accentuates the common overall liberal-conservative dimension seen going from southwest to northeast. This is because issue-by-issue variation for an individual tends to get averaged away, leaving a stronger indication of the "ideology" part that is correlated across an individual's responses.

Figure 13: Ideological constraint as polarization?



and other things equal). If politics in the black-dot society happens to lead to moderation on both dimensions – the green dot, which is at the mean and median on both x_1 and x_2 – average unhappiness will be much lower than for the same outcome in the more diverse gray-dot society. "Polarization" in this sense that is associated with ideological constraint is only problematic to the extent that it occurs along with increased variance or is accompanied by political bias away from moderate policies.

22.3 Why is U.S. political competition approximately one dimensional?

A repeated and strong finding in research on public opinion in the U.S. is that most citizens exhibit very low levels of ideological constraint (the classic reference being Converse, 2006, first published in 1964). From the perspective of political scientists and others who spend a lot of time and energy following politics, they appear to be all over the place, or "innocent and confused" as Kinder and Kalmoe (2017) put it.[106]

In sharp contrast, elected officials in the U.S. have been shown to present high levels of ideological constraint, particularly as expressed in how they vote on bills in legislatures. Poole and Rosenthal (1985, 1991, 2000) showed that a one-dimensional spatial model could classify more than 80% of individual

[106]The political science literature in this area is rich with condescension towards most voters, largely on the grounds that they take "liberal" positions on some issues and "conservative" positions on others, and that they often misattribute the liberal/conservative label relative to what party positions are. Because it equates the party packaging of issues with "ideology," I am not sure this condescension is well merited. There is a related tradition of also a tradition Bin the Political ehavior of concern that citizens are simply following political elite's

House and Senate member roll-call votes for almost all Congresses up to 1995. In their 2000 book, they found that starting in the mid 1970s, the one-dimensional model fit increasingly well, classifying about 90% of roll-call votes correctly by 1995.[107] Poole and Rosenthal (2001) finds that a one-dimensional model similarly classifies upwards of 85% of roll-call votes for a set of other legislatures, including the European Parliament, the U.N. General Assembly, and sessions of the French, Czech, and Polish parliaments.

At least for the U.S. case, then, the situation appears much as depicted in the left panel of Figure 13, with public opinion distributed like the gray-dot society and Congress tightly constrained on approximately one dimension like the black-dot society (and in reality with far more than two issue dimensions for public opinion). Why would this be?

Perhaps some of the explanation is that "ideology" is form of political knowledge, and as political elites and professionals, members of Congress are much more politically knowledgeable than the most voters, who have no reason to pay close attention. Poole and Rosenthal (2001) characterize Converse's concept of ideology – which underlies this whole literature in American Politics – as "the knowledge of what goes with what." The politically well-informed person in the U.S. knows, for example, that if you believe that taxes should be lower, you "should" also be opposed to raising the minimum wage. If you are also opposed to abortion, you will be demonstrating even more "ideological constraint," even though there is no necessary or direct logical connection between any of these policy preferences. Perhaps it is that politicians and political junkies like political scientists are acutely aware of, and adopt as their own preferences, how political parties happen to bundle preferred issue positions. This is what we then come to define as "ideological coherence" or "constraint."[108]

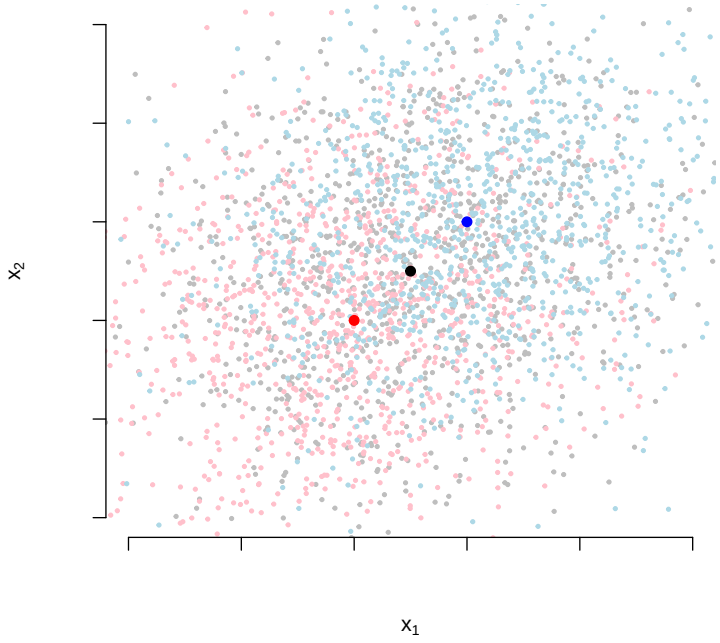
But here is another, not mutually exclusive, possibility. Think of the black dots in Figure 13 as the ideal points of members of Congress, who represent particular geographically defined constituencies, each with voters who have diverse preferences on the two issue dimensions. Figure 14 depicts hypothetical voter ideal points in three districts, indicated by light red, gray, and light blue dots. The dark red, black, and darker blue dots show the means (and median) positions for each district. The fact that the means fall on a line with increasing slope reflects that in the population as a whole there is a small positive correlation between voters' preferences on the two dimensions. In this example, there is almost zero correlation between voters' ideal points on the two dimensions *within* each district. So in this example voters *within* a district have almost zero ideological constraint, and they have very low ideological constraint when considering society as a whole. Across districts, however, the *average* voter is somewhat more conservative or more liberal on both dimensions. This is actual empirical pattern in U.S. House districts, for example.

If political competition tends to favor and select candidates who adopt positions close to the issue-by-issue mean or median within districts, *then in the legislature we will get a set of politicians who exhibit very high "ideological constraint"!* A centripetal force like that identified in the baseline Downsian model would turn slight correlations across issue preferences in the population into a tight relationship

[107] what now?

[108] Huber and Powell (1994, 294) hypothesize that near-unidimensionality may be the result of political competition, with politicians driven to make things "intelligible" for cognitively-constrained voters: "Indeed, it may be that the ability of students of legislative voting behavior to describe voting over long periods of time in a single dimension and the ability of students of electoral behavior to describe party competition in many different countries using a single dimension reflect the need for democratic debate to reduce conflict to a single dimension in order to make it intelligible."

Figure 14: Median voter-seeking favors "ideological constraint" in a legislature



in the legislature, well represented by a single "left-right" dimension.

This argument works the same way for more than two issues: Dimension reduction is a strategic consequence of median voter-seeking dynamics in electoral competition, in the presence of some positive correlation in issue means across districts. (That is, districts that are more liberal on some issues tend on average to be more liberal on other issues, even though the within-district, cross-issue correlations can be very small or even non-existent, so that voters show very low ideological constraint both within districts and in the society as a whole.)[109]

It is true that we have not yet considered two-candidate electoral competition with more than one issue

[109] Suppose we have $n = 435$ districts, and each has a large number of voters who have single peaked preferences over $k > 1$ issue dimensions, each one represented with the real line as usual. In district j, voter i's ideal point on issue k is x_{ijk}. Suppose that in each district j, voter preferences over the k issue dimensions are drawn from a multivariate Normal distribution such that there is zero correlation across the k issue dimensions. Voters within districts thus exhibit *zero* ideological constraint in Converse's sense.

However, suppose also that there is some correlation in the issue means across districts. For instance, let $\bar{x}_{jk}$ be the mean on issue k in district j, and suppose that the set of k issue means for each district j is drawn from a multivariate Normal distribution that has some positive correlation in the bivariate correlations across means. Thus, if a district's voters are slightly more liberal than other districts on issue $k = 1$, they are also likely to be slightly more liberal than average on the other issues. Still, there is no correlation within the district in people's liberal/conservative positions across issues. So there are more liberal and more conservative districts, where this refers to average opinion on each issue. But there is zero ideological constraint within districts and almost zero constraint if we measured the whole population (if most of the variation is at individual rather than district mean level).

Then median/mean representatives will tend to fall on line in the k-dimensional space.

dimension. You might wonder if the centripetal tendency seen with one dimension still obtains with multiple dimensions. In Sections X below we will see that under reasonable assumptions, it does.

22.4 Representation and democratic performance. Deliberation.

In the U.S. and almost all other countries, members of the national legislature represent geographically defined constituencies. Returning to Figure 14, imagine that it depicts the ideal points of voters in three geographically defined districts of a country whose national legislature has only three members. If each politician advocates for the indicated district median position, it would be hard to disagree with the claim that this is "good representation" *at the level of the district*.[110] But what about at the national level, once these representatives go to the national legislature and choose policies there? For that matter, what exactly does, or should, "good representation" even *mean* at the national level?

On a little thought, it is immediately apparent that for a legislature, "good representation" of voters' preferences and good democratic performance are not the same thing. We could easily have good representation but bad performance, if (say) legislative politics produces policies far from the national mean or median despite having a highly representative legislature. And we could have good performance but poor representation. For example, legislative politics might produces policies close to the national mean or median of citizens' preferences, even though typical policy-bundle preferences in the population are not well represented among the set of elected members. In Figure 13, for instance, median-voter-theorem legislative politics would produce an outcome close to the national median/mean, but, as noted, libertarians and socially conservative economic liberals are poorly represented in the legislature as a whole.

I should make explicit the implicit standard for "good representation" adopted in the last two paragraphs. It holds that representation of a district or in a legislature is better, the smaller the average distance between the policy preferences of each voter and his or her closest representative (in terms of policy positions). At the district level – and assuming single-member districts – this just says that good representation is when the elected politician's positions are the same as those of the typical voter. But for a legislature with $n > 1$ members, optimal representation involves choosing n points in the policy space so as to minimize the average distance of voter ideal points in the population to the closest politician.

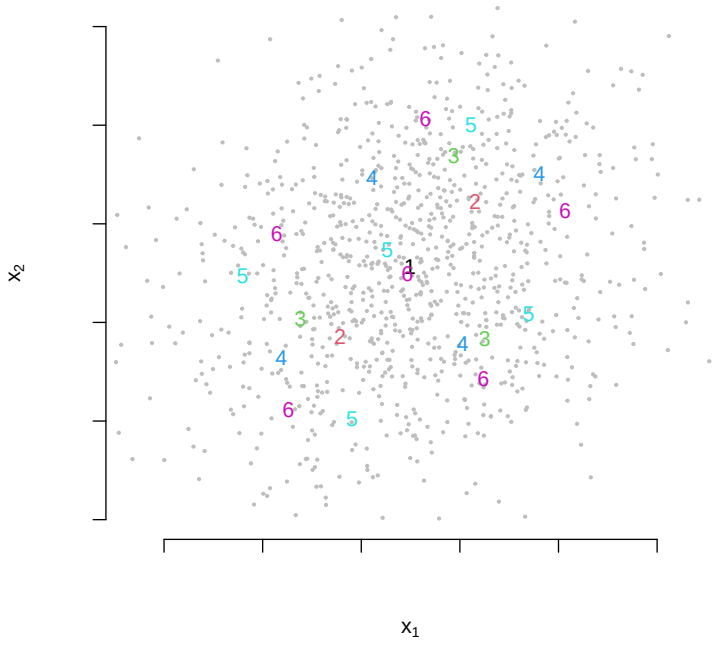
To illustrate, Figure 15 displays roughly optimal placement of legislator positions for a legislature with one to six members, given the voter ideal points indicated by the gray dots.[111] Rather obviously, increasing the size of the legislature will, with optimal or near-optimal "representation" in this sense, reduce the average distance to one's closest match in terms of policy preferences.

"Representation" is plausibly better in the six-member legislature shown Figure 15 than in the one-member legislature. But so what? If we are linking the idea of good democratic performance with how close policy *choices* are to what the typical citizen would like, then why be concerned about

[110] Putting to the side the matter of what information the public in each district is basing their preferences on.

[111] So for example the "1" shows the global mean, which would be optimal representation (and optimal performance on policy outcome) in a "legislature" with just one member. The "2"'s show optimal placement to minimize average distance in a legislature with two members; and so on. The positions were found using k-means clustering.

Figure 15: Optimal representation for legislatures with one to six members



"representation" in the above sense? If the one-member legislature (the President?) would choose policies at the national means or medians, why isn't this just as good as or better than a large legislature that represents the diversity of policy preferences in the population, but which may or may not settle on policy choices close to what the typical citizen would like?

One plausible response is that the democratic politics of legislatures is not just about aggregating fixed or well-considered public issue preferences. It is also about *identifying and implementing* good policies in the face of huge uncertainties, both about the relative importance of different issues to different groups, and about tradeoffs across issues and technical or practical problems concerning what works or is feasible (recall sections 21.3 and 21.4).

If so, then better representation in a larger legislature may help with the identification and implementation of policy choices that better reflect (a) actual mean positions on diverse issues, perhaps especially those with significant mean-median conflicts, and (b) policy choices that better reflect what voters would want if they were better informed. Having a single representative, or a legislature filled with politicians advocating for a narrow set of policy bundles relative to the whole space, might miss out on arguments, perspectives, and relevant information from, for example, farmers, libertarians, socially conservative economic liberals, racial and ethnic minorities, and so on.

Imagine a legislature composed entirely of representatives of the typical voter in the country. Now imagine adding a representative who understands and advocates for the perspectives and policy preferences of some non-typical, or minority, constituency. It could easily happen that interaction in the

legislature now allows for the identification of ways that policies could be adjusted so as to improve things for the minority at little or no cost – or even some benefit – for the majority.[112]

Various empirical, positive, and normative literatures in political science refer to information sharing and exchange of reasons to reach better or more consensual decisions as *deliberation*.[113]

Game theoretic methods as they currently exist provide some means to formally represent and analyze one aspect of deliberation – deliberation as information sharing. The simple formal framework that I have been using in this section to represent policy preferences on multiple dimensions is not up to the task. But it is possible to extend it in this direction. For example, we can introduce legislator uncertainty about what is the voters' most preferred policy bundle, and have this be discoverable to some extent by investigative effort, whether via technical policy analysis or communication of different "signals" observed by different legislators and/or lobbying groups.

The exercise below takes you through a simple illustration, in which a legislator representing an intense minority position on a specific issue also has better information than the median legislator about likely effects of different policy choices on this dimension. Depending on the extent of the mean-median conflict, and also the quality of the minority advocate's information, "deliberation" in the rather reductive sense of simple information transmission may be able to make both better off. This illustrates a defensible argument for why good representation in a legislature, in the above sense, can be valuable for good democratic performance.

But information transmission and discovery is arguably a rather cramped idea of deliberation. It doesn't adequately capture persuasion and opinion change due to exchange of reasons, or moral argument.

Whether deliberation is understood as information sharing or moral argument, if the effect is to increase agreement on what are good policies, then deliberation has a welfare consequence worth highlighting. Recall the decomposition based on quadratic utility and a policy space: Average welfare loss = policy bias + preference heterogeneity. Trying to reduce policy bias by design or reform of political institutions is one thing, but what can be done about preference heterogeneity?

Ideally, deliberation reduces levels of disagreement by moving participants' views of what would be best towards each other, so reducing variance. My impression is that this is not an important element of how political theorists want to justify "deliberative democracy." Their hope is that it would lead to actual moral improvement in policies chosen, or policies that are more rightly chosen. But from a welfarist perspective, the disagreement-reducing consequences could nonetheless be important. It could be that in some democracies overall, or on particular issues, policy bias is minimal but democratic performance (in the sense of average satisfaction) is still very bad because so many people disagree strongly with the policies favored by the typical voter. How else to improve wellbeing except by some form of

[112] A well-run civil service in government agencies or ministries plays some of this informational and analytical role, but having political principals in the legislature motivated to look out for diverse constituencies and interests is probably an important for motivation and oversight of the bureaucracy. In an argument against geographically defined constituencies, Rehfeld (2005) proposes that a better alternative would be to assign each citizen at random and for life to a legislative constituency, so that every member would represent a random subset of the population. With median-voter-seeking dynamics, this would produce a legislature of very similar members, all fairly close to the typical voter in the population as a whole, but not representing the diversity of views in the population well at all.

[113] cites. The literature on "deliberative democracy" puts additional conditions on "exchange of reasons" in defining what should count as deliberation (e.g., Cohen 2005).

deliberation, and isn't the form and effectiveness of deliberation amenable to institutional design and policy choice as well? Policies that bear on the structure and operation of public and private media are an example.[114]

A simple example of deliberation as information transmission. The median legislator – call her M – will choose a policy $x \in \mathbb{R}$, on an issue dimension that is of special concern and impact for a minority of other legislators (the "minority," or m). Policy preferences for both sides depend on facts about the world that are better known by the minority than by M. But it is also known that the minority prefers a more conservative policy than M for any given "state of the world." We will ask when the minority can credibly convey their "private information" about the effects of the policy to M.

Suppose that preferences over the policy depend on two things – ideological, or distributional, implications on the one hand, and efficiency considerations (cost, or how well it will work), on the other. Putting the cost/efficiency considerations to the side, M's ideological ideal point is zero, with quadratic-loss preferences $-x^2$. m's ideological ideal point is the conservative policy $\delta > 0$, with quadratic loss for x of $-\alpha(\delta - x)^2$, $\alpha > 0$. You could imagine that the policy $x = \delta$ is a tax break, or regulatory concession, to an industry that is concentrated in the districts represented by the legislators in the minority group m.

Maybe a tax break or regulatory concession for this industry would have positive aggregate economic effects that would partly compensate the majority for the loss of tax revenue or the harms of the deregulation. Call that "state of the world" $\omega = \delta$. Alternatively, maybe the tax break or regulatory concession would cause harms and losses beyond those captured in the ideological/distributional quadratic loss terms, and which would affect both M and m in the same way. Call this state of the world $\omega = 0$. Suppose that the parties' preferences are both affected by a common cost/efficiency term $-\beta|\omega - x|$, which means that for this dimension, they both would ideally like to match the chosen policy x to the state of the world $\omega \in \{0, \delta\}$.

Putting these two dimensions of preference together, we have the following objectives for M and m:

$$u_M(x) = -x^2 - \beta|\omega - x|, \text{ and}$$
$$u_m(x) = -\alpha(\delta - x)^2 - \beta|\omega - x|.$$

In short, they have opposed ideological or distributional preferences over the policy, but also a common interest in adjusting the policy in the direction of the state of the world ω (that is, towards $x = 0$ when $\omega = 0$ and towards $x = \delta$ when $\omega = \delta$). The parameter $\beta > 0$ measures the degree of common interest, the importance of the unknown facts about the policy's effects relative to the distributional or ideological concerns. The parameter $\delta > 0$ measures the degree of preference disagreement between M and m on the ideological/distribution dimension, and the parameter $\alpha > 0$ measures the minority's intensity of preference for departures from their preferred policy $x = \delta$.[115]

Exercise 32. *(a) Suppose that the state of the world $\omega \in \{0, \delta\}$ is commonly known by both M and m. Work out each player's most preferred policy in each state of the world. If we take the "degree of polarization" to be the distance between their preferred policies (conditional on knowing ω), what is this in each state of the world?*

[114] ... [also the problem of education policies]

[115] δ and α capture the two forms of polarization discussed in footnote 2.

(b) Now suppose neither player knows ω but both think it is equally likely to be 0 or δ. What are their most preferred policies in this case, and what is the (naive utilitarian) sum of their utilities if M chooses the policy? What is the ex ante expected degree of polarization in this case (that is, distance between preferred policies given that both think $Pr(\omega = 0) = 1/2$?

(c) Next, suppose that the minority observes the state of the world, while M does not. M thinks it is equally likely that $\omega = 0$ or $\omega = \delta$ and m knows this. Find conditions on δ, α, and β such that m would want to truthfully report ω to M, expecting that M would then implement M's most preferred policy based on m's report. Interpret these conditions, and then compare the ex ante expected sum of the players' utilities when truthful information transmission is possible to the sum in (b), where M was just choosing based on no information.

(d) Suppose that the condition in (c) does not hold, so that m would not be willing to report truthfully in all states of the world. But imagine that M can commit ahead of time to implement a policy different from her ideal point, conditional on m's report. For example, consider the specific case of M delegating the choice of policy x to m. Under what conditions would M want to do this ex ante? Explain why this would make sense for m in ordinary language.[116]

22.5 Representation and democratic performance. Geographic districts.

So we have some reason to think that a more representative legislature has the capacity to perform better. Better information revelation (stemming from more diverse perspectives and knowledge) has the potential to reduce policy bias, the distance of chosen policies from typical voters' most preferred outcome. At the same time, better information revelation can increase average satisfaction by reducing the average extent of disagreement with any given policy choice (preference heterogeneity).

Let's return then to Figure 14, which illustrates how optimal representation for a set of geographic districts could make for not-so-good representation of preference diversity in the country as a whole. What can be done?

Rather obviously, the problem there is closely linked to having *geographic districts* as the basis for representation in the legislature. Unless the political preferences of the typical voter in every (geographic) district are like a random sample of the entire population, then with single-member districts *there is an inevitable tradeoff between good representation at the level of districts and good representation at the national level*.

As the example further illustrates, if the preferences of typical voters across geographic districts are correlated across issues – so that, for instance, being more "conservative" on one issue associates with a higher probability that the typical voter is more conservative on other issues – then the legislature will tend to represent a single dimension of political competition no matter how many dimensions there are in the population's preferences![117]

[116]Notes on this example: First, it is possible to set up and analyze this problem as a proper extensive-form game, using formal tools to be presented in Part IV. Second, Krehbiel (1992) argues that median legislators in the U.S. Congress use agenda powers they confer on congressional committees to delegate some power to choose policies away from their first preference in order to incentivize information revelation that improves the policy in other, commonly valued ways.

[117]Recall that in the Converse tradition, what makes a direction on an issue "liberal" or "conservative" is which party

Thus the question: Why use geographic constituencies to structure representation in legislatures?

U.S. citizens may be puzzled, wondering just what is the alternative to geographically-defined constituencies. In fact there are many other possibilities, some of which are useful to consider by way of contrast.

- *Multimember districts*. The Israeli Knesset has 120 members, elected from a single national "district" (the whole country). Voters vote for parties rather than individuals in a "party list" system. Parties are then allocated seats approximately proportional to their vote share, if they receive enough votes to get above a threshold (currently 3.25%).[118]

  This relatively pure *proportional representation* (PR) electoral system addresses the representation problem of single-member, geographically-defined constituencies identified above. Consider Figure 14 again. With PR and a single national district, you can see that there would be an incentive for *entry* by new parties seeking to represent voters with preferences, say, to the northwest or southeast of the black dot (the median in all directions). In fact, with a low threshold, you could have many parties entering to represent voter preferences clustered in many parts of the policy space. Recall that in reality, there may be many more than two issues dimensions of concern, and many "single-issue" voters.[119]

  So, on grounds of good *representation* of the broad spectrum of policy preferences in a large population, a PR system seems attractive relative to plurality rule with single-member districts. An additional attractive feature is that PR with party lists lets electoral competition determine the structure of political cleavages. It does not prejudge and build in by design any particular way of grouping interests, whether by geographic boundaries or some set of individual characteristics, as in the consociational systems mentioned below.

  What's the catch? Pure PR with party lists might be weaker when it comes to *accountability* (Huber and Powell, 1994; Powell, 2000). One votes for a whole party. Who is *my* representative? If I have a problem with a national government service in my local area, or with government policy as it affects my area, who has an incentive to advocate for me and others in my local community?

  Further, when the government is run by a coalition of parties that form a majority in the legislature, it can be difficult to evaluate the performance of any one party. Accountability for performance attaches to majority coalitions rather than to the specific parties (and individuals) that voters choose between in the election.

  There are many ways of combining multimember districts with geographical representation in the design of electoral systems and legislatures. For example, in the German Bundestag, half of the 598 seats are elected from single-member, geographic districts by plurality rule, and the other half by a PR, party-list system at the level of German states. So an individual voter has both a

espouses it, not substantive features of the issue. So we are free to choose the meaning of "liberal" and "conservative" directions on each issue however we please – for example, by calling the positions associated with one party "liberal" and the positions associated with the other "conservative," whatever they are.

[118] Small parties are allowed to form coalitions to get above the threshold, so in fact even smaller parties can gain representation.

[119] Section X below briefly considers results for the Downsian model when we allow for more than two parties or candidates, and for electoral systems other than plurality rule, like PR. But this is a large area and it is not well covered in this book. See Cox (1997) for references and a classic treatment.

local representative and, if she votes for a party that wins seats, representation by her preferred party.

- *Consociational systems*. It is also possible to form electoral constituencies out of *types of people*. Pre-Revolutionary France had an institution known as the Estates General, which represented three classes – clergy, nobility, and commoners – in three different assemblies that met concurrently (on occasion). Lebanon's electoral law employs party-list PR but assigns quotas for seats in the parliament to 11 religious groups (7 Christian sects, 4 Islamic). Many countries have introduced legislative quotas for women. One can imagine any number of other ways that one could, in principle, divide up society into groups that would elect representatives to the legislature.

  This approach fixes some political cleavages outside of regular electoral politics, creating incentives for politicians to cultivate and reinforce particular political and social identities. This is also true of geographic constituencies, although these are generally more permeable than ascriptive social identities (to the extent that people can move).

If single-member electoral districts may have an advantage for accountability,[120] this doesn't yet imply that the *geographical* constituencies would be the best, or a good, way to define them. Why do geographical electoral districts seem so natural and reasonable to us? Is it just tradition?

As a general matter, with single-member districts for a legislature, two considerations favor grouping voters by *degrees of common interest*. First off, as we have already seen in sections 20 and 22.1, more similar preferences over policies increases average contentment with any one choice. Average dissatisfaction with a good representative will be lower in a more politically homogeneous community.

Second, accountability considerations probably also favor grouping voters by common interests. To see why, consider a district with voters divided roughly equally between extreme conservatives and extreme liberals, and a representative who happens to be from the more liberal party. The conservative voters are disinclined to vote for reelection regardless of the representative's performance on any shared interests, and the liberal voters are disinclined to punish poor performance on shared interests if this means getting a conservative party replacement. Liberals in the district thus have a stronger incentive to overlook corruption or incompetence than if the policy positions of a challenger were not so different.[121]

A reasonable argument – or start of an argument – in favor of geographic districts is that people who live near to each other tend to have many important interests in common, just because they live near to each other. For example:

- A large number of the most valued government services (including services that require significant regulation) are spatially distributed, such as schools, roads, ports, electrical and gas grids, and water and sanitation infrastructure. People who live near each other, particularly in cities and towns, tend to have strong common interests in the provision of spatially distributed public goods. They also often have more common views on the types and levels of such public goods, as compared to people in other locations.

[120] I am not sure what the empirical evidence says about this. Powell ...

[121] We consider a simple model of this accountability problem in section X below. See Padró i Miquel (2007) and Alesina, Baqir and Easterly (1999) for more extended treatments.

- Different industries often co-locate due to natural resources or features like farmland, coal, waterways, ports, weather (e.g., Hollywood, Napa), or due to scale and "agglomeration" economies (e.g., Silicon Valley). With many people working in the same industry and with jobs in businesses that depend on the local industry, common interests over policies can follow.
- People who live in a particular area often have common cultural, religious, or historical traditions that bear on their preferences over public policies, both locally and nationally. This may have been more true in the past, but is still often the case.

Summing up, single-member, geographically-defined electoral districts may have advantages for accountability and average satisfaction with representatives at the level of the district. But there is a tradeoff: Good representation at the district level works against good representation at the national level, as illustrated by Figure 14 (compare to Figure 15). By the arguments in the previous section about deliberation, approximately one-dimensional representation in the legislature could make for worse outcomes in terms of both bias and preference heterogeneity, relative to a legislature that better represents diverse positions and the information behind them.

22.6 Aggregation problems. Malapportionment and maldistribution.

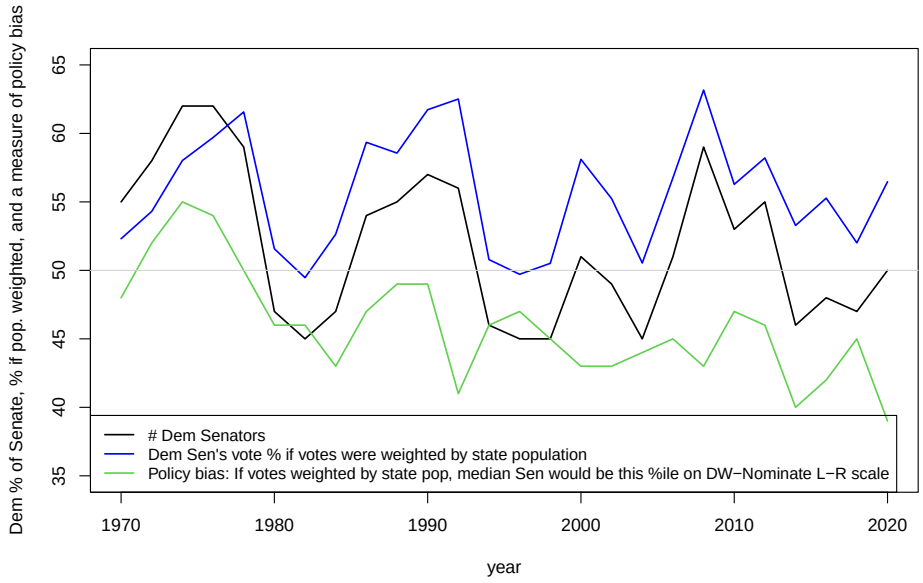
Supposing a commitment to single-member, geographically-defined electoral districts, how should these be drawn for the sake of good democratic performance? This is a difficult and multifaceted question that deserves more attention – and more clarity and progress than you will find here.

Malapportionment. On first thought it would appear incontestable that the districts should be equally sized in terms of population. If not, then "one person, one vote" is violated, which many political theorists see as bad independent of consequences, because it violates the valued principle of political equality. From our naive utilitarian perspective, malapportionment has the further demerit that violating "one person, one vote" inclines to democratic failure in the form of policy bias, to the extent that the preferences of the over-represented are systematically different from those of the under-represented.

The US Senate is a notorious example. Figure 16 shows the percentage of Democratic Senators for 1970-2020, along with what the percentage of Democratic Senators' votes in the Senate would be if votes were weighted by state population. Starting in the late 1970s Republicans increasingly came from smaller states relative to Democrats, making for a large degree of Republican over-representation. With votes apportioned by population (or equivalently, equal-sized states and the same set of Senators from each), Democrats would have held a Senate majority *in 24 of these 26 Congresses*, versus only 15 under malapportionment.

As regards bias away from the national mean or median in the Senate, this is a problem only if Republican and Democratic Senators elected from the same state would take different positions. They do, of course, a failure of Downsian convergence that we investigate at length in Part IV. But even if convergence was complete on national policy issues, the US federal government distributes a lot of resources that support spatial public goods. As discussed above, for such things Republican and Democratic Senators from the same state tend have a common preference for more stuff. If any Senator from Wyoming is going to vote for distribution of federal largesse to Wyoming, then Wyoming voters

Figure 16: Republican under- then over-representation in the Senate,
and a measure of policy bias due to malapportionment



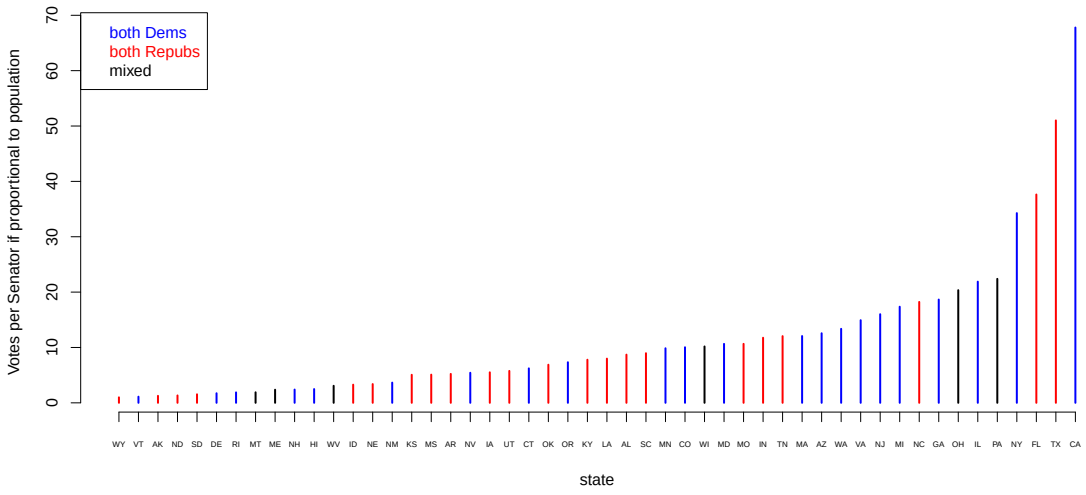
of both parties get roughly *70 times* more electoral clout on such matters in the Senate than do voters of both parties in California. A good-sized empirical literature finds evidence that over-representation due to malapportionment in legislative bodies associates with greater per-capita transfers and spending on spatially distributed goods.[122]

For national policy issues, we can use estimates of Senators' Left-Right positions based on their roll-call voting records to give a sense for the degree of policy bias related to malapportionment. The green line in Figure 16 uses the DW-Nominate scaling of the Left-Right positions of Senators (based on their roll-call voting records) to show how much more liberal the median (50th) Senator would be if Senate votes were weighted by state population. For example, it shows that for the Senate elected in November 2020, if votes were weighted by population then the median Senate vote would be that of John Hickenlooper (D) of Colorado. Hickenlooper is estimated to be 39th most liberal Senator by DW-Nominate – as compared to the one-Senator, one-vote median, who is Joe Manchin (D) of West Virginia.

Can there be any warrant for drawing geographical electoral districts to have different population sizes? Indeed there can. Just above, I argued that both electoral accountability and average voter satisfaction with their elected representatives (lower σ^2 among the voters) may be increased by drawing boundaries around communities that share more common interests or views (see also sections 20 and 22.1). This is likely to hold with special force for areas with natural economies of scale for the provision of public goods, cities in particular. Since geographic communities of relatively similar interests will not in general have equal-sized populations, the principle of "one person, one vote" – desirable for minimizing policy bias – could well be at odds with these other desirable features of drawing electoral district

[122] See especially Ansolabehere, Gerber and Snyder (2002); Snyder and Ansolabehere (2008). ...

Figure 17: # Votes per Senator if proportional to population, and party composition after the 2020 election



boundaries around "natural communities."

Imagine, for example, a state large enough to have two House districts, but with most of its population concentrated in two cities. The smaller city is very conservative and the larger very liberal. To have equal-sized, geographically compact districts one might then need to have a minority of liberal voters who lived nearer to the big city districted with the conservative city, to the unhappiness of these liberals and possibly the conservatives as well. Average satisfaction with a mean or median representative in the more conservative district would be greater if unequally-sized districts were permitted – although the larger city's district would then suffer some effects of adverse malapportionment.[123]

But in fact there does not need to be any tradeoff here! There is another way to achieve "one person, one vote" without requiring that electoral districts have equal-sized populations. Just weight the legislative votes of the representatives by the relative size of their districts. In the U.S. Senate, for example, give Texas Senators 50 times more votes than the Wyoming Senators, California Senators one-third more than Texas Senators, and so on. Such schemes are not unheard of – in fact they are fairly common in international organizations – but so far as I know they are rarely if ever used in national constitutions. I don't know why not, although I suppose that in the case of federalisms, if union required unanimity to begin with then this may have "baked in" malapportionment at the federal level (thinking of the US example).[124]

Figure 17 displays the number of votes Senators from each U.S. state would get if "one person, one vote" were implemented in this manner, using 2020 population data and the 2021 partisan composition.

[123] In addition, the liberals in the vicinity of the big city who are districted with the small city do not have a representative directly motivated to care about their public goods provision in the big city. In practice, many states' laws or constitutions bid redistricting bodies to try to keep municipal and other "natural" communities boundaries intact in state-level legislatures (Rodden, 2019, 140, example of Pennsylvania).

[124] Snyder, Ting and Ansolabehere (2005) analyze the strategic logic of legislative bargaining under weighted voting rules.

Maldistribution? Policy bias in a legislature can also arise from the way that voters with different political preferences are distributed across districts. For instance, if more liberal voters are concentrated in a minority of districts whereas the rest have closer majorities of more conservative voters, then the median of the district medians can be more conservative than the position of the national median voter. Rodden (2019) argues that this pattern characterizes the electoral geography of many advanced industrial democracies. Urban areas inefficiently concentrate liberal votes in a small number of districts, which helps to explain "why cities lose."

Consider dividing an area that has 650 liberal Democrats and 350 conservative Republicans into three roughly equal-sized districts. If one could district the 350 evenly between two districts – 175 and 175 – together with 160 and 160 of the Democrats, this leaves one district with 330 Democrats. Then conservative Republican candidates win the two more competitive districts and the median legislator is a conservative Republican, even though liberal Democrats form a solid 65% majority of the whole area. Call this Scheme R.

Fix all of the Democrats' ideal points at $x = 0$, the Republicans' at $x = 1$, and assume linear or quadratic preferences on $[0, 1]$. Then Scheme R implies an average policy loss of .65 if the median legislator rules, choosing $x = 1$. With quadratic preferences the optimal policy would be the mean, $x = .35$, yielding an average policy loss of $.65(.35)^2 + .35(.65)^2 = .22$. If districts were drawn to put liberal majorities in each one – Scheme D – then all three legislators are Democrats and the policy outcome is $x = 0$, for an average policy loss of $.35 < .65$.

Notice that Scheme D is not "obviously correct" or optimal. An alternative scheme with two almost completely Democratic districts and one completely Republican district would still have the median legislator as a Democrat. But at least the Republican voters are represented in the legislature (good for deliberation, inter alia) and all voters are maximally close to their representatives. Call this Scheme PR.

Schemes R and D are called gerrymandering if they are the result of a process in which one party's members or advocates deliberately draw the lines for the sake of partisan advantage in the legislature. By our naive utilitarian criterion, gerrymandering seems to be clearly a bad thing – a political failure – in that it constructs policy bias away from the average or median voter's preferences by manipulating preference aggregation given single-member geographic districts.

But not so fast. Gerrymandering for the sake of partisan advantage is clearly bad. But as argued at the start of this section and as evident from consideration of Scheme PR, policy bias in the legislature is not necessarily the only relevant normative consideration for evaluating democratic performance. What about average satisfaction with one's representative, which is higher in more politically homogeneous districts?

Notice that the gerrymandered Scheme R has an advantage over Scheme D in this example in terms of average distance of voters to their representative. In Scheme R this is roughly $(0 + 1/2 + 1/2)/3 = 1/3$. This compares favorably to Scheme D, which has about 650/3 Democrats and 350/3 Republicans in each district, for an average distance of .65. Scheme PR is best on this score, with almost all voters in full accord with their representative.

Should we care about average distance of voters to their representatives when evaluating districting

schemes? At the national level, policy loss from preference heterogeneity depends on distance from the policies chosen by the legislature, not distance from one's own representative. But many believe that there is something to be said for having "one's own" representative in the legislature.

Indeed, in *Thornburg v. Gingles* (1986), the Supreme Court ruled that district lines drawn in such a way as to "dilute" the votes of African Americans could be, if other conditions obtained, a violation of the 1965 Voting Rights Act and the Constitutional principle of one person, one vote. This ruling effectively endorsed the creation of "majority-minority" Congressional districts that concentrate racial minorities, making own-race representation much more likely. However, because districting African-Americans in smaller numbers across more competitive districts could increase conservative candidates' success, majority-minority districting could in some situations have the "perverse effect" of reducing legislative representation of policy positions typically favored by African Americans, even as it increased descriptive representation.[125] From the normative perspective developed above, majority-minority districts can have the value of reducing the average distance of voters to their representative's positions.

This is simply a potential tradeoff that appears in a system that has equal-sized, single-member, plurality-rule electoral districts. Gerrymandering: Bad for policy bias if done to advantage a party in the legislature. Forming districts around communities of interest with similar political preferences: Can be good, both for satisfaction with representatives and perhaps accountability.[126]

With multimember districts and proportional representation, one doesn't draw lines (or as many), and as we saw above (Figure 15) the prospects for lower average distance between a voter and at least one party in the legislature are probably higher. But as noted there may be downsides for accountability and good representation of small natural geographic communities of common interest.

Just how significant are the negative welfare effects of maldistribution in U.S. national politics? Rodden (2019) argues that they are a powerful source of electoral disadvantage for the Democratic Party, in terms of seats won in state and the federal legislatures.

How much this matters for policy bias, however, depends on another factor that we should be clearer about. Do elections and legislative politics tend to select median or mean candidates and policies?

> **Observation**. Suppose that electoral politics in (equal-sized) districts produces winners who adopt roughly the *mean* policy positions of voters in their districts, and that legislative politics likewise select the *mean* position of the representatives as the policy choice. Then the policy outcome is the same *regardless of how district boundaries are drawn*. This is because regardless of how voters are assigned to districts, the mean of the district means will be the same. By contrast, the median of the district medians, or means, can depend on district boundaries.

[125] Shotts (2001) observes that introducing majority-minority districts can increase "substantive representation" (favoring less policy bias in the legislature) in some situations, and decrease it in others. There is an empirical literature on whether or how large the "perverse effect" is. ...

[126] The term gerrymandering is also often used to cover cases where district boundaries are drawn to protect incumbents by giving them substantial partisan majorities. This kind of gerrymandering can work *against* partisan advantage, even as it would tend to increase political homogeneity and so favor reducing the average distance of voters from their representatives. Conventional wisdom sees this kind of gerrymandering as (also) bad, because it reduces competition across parties. However, if there are open primaries, it is not clear to me why there is necessarily a problem here.

Majority rule in principle favors median positions over means, providing warrant for the concerns about maldistribution as a source of democratic failure. Still, as we will discuss at some length in Part IV, the stronger motivations of people and groups farther from the median can tend to work to pull policy towards means when means and medians differ. To the extent that this is so in both electoral and legislative politics, the effect of maldistribution on policy bias attenuates.[127]

Exercise 33. *Figure 18 is a very stylized depiction of a circular country that has a circular urban area in its center. Every voter in the urban area is a liberal, and every voter outside the urban area is a conservative. Voters are uniformly distributed in the territory, and share $l > .5$ of the population are (urban) liberals while r are (more rural) conservatives ($l + r = 1$). Suppose the legislature has n single-member districts, n odd. For doing the exercise, start by working out the case of $n = 5$ and then generalize to n.*

1. *Draw district boundaries to form five equal-sized districts (thus $1/5$) that all have a majority of liberals, so that the legislature would be all liberal representatives.*
2. *Draw district boundaries to form five equal-sized districts that return three conservatives and two liberals to the legislature, with the conservative-liberal vote shares equal within the three conservative districts. Assuming your solution has two districts that are entirely liberal, what is the condition on l necessary for this to be possible?*
3. *Let liberals all have policy ideal points at $x = 0$ and conservatives at $x = 1$ in the policy space $[0, 1]$, with quadratic preferences. For both (1) and (2) compute:*
   (a) *Average policy loss in the legislature if the national policy is the median of the representatives positions, which are the medians of their district voters' preferences.*
   (b) *Average policy loss in the legislature if the national policy is the mean of the representatives' positions, which are the means of their district voters' preferences.*
   (c) *Average distance from one's representative.*

[127] Policy bias related to malapportionment of the Senate is likely much larger than bias related to maldistribution in the House. This table shows the results of regressing Republican seat share (minus .5) on Republican national vote share (minus .5) for both Senate and House elections for each Congress from 1976 to 2020. The intercept is thus an estimate of the size of the "bonus" seat share in each chamber that Republicans get because of malapportionment and maldistribution, respectively. (The standard errors of these estimates are extremely small, except for the House intercept which has $p = .03$.) Five and a half seats in the Senate (out of 100), about 7.4 seats in the House (out of 435).

	Senate	House
Intercept: Est. of Rep seat share bonus at 50-50 national vote	0.055	0.017
Est. seat share gain for 1% vote share gain	0.011	0.018

Figure 18: Urban core and more rural periphery

Part IV

Policy bias in US electoral politics

23 Downsian centripetal forces?

We have seen that in the two-candidate Downsian model with a single policy dimension, candidates choose policies at the position of the median voter and the race is very close – each is equally likely to win.[128]

It is clear that candidates in US electoral contests do *not* normally adopt and espouse exactly the same policy positions. It is also obviously true that for Senate and House races, only a minority are close. I will go over some basic evidence on "divergence" and "safe seats" in the next section.

At the same time, however, there is also plenty of evidence that the strategic pressures and incentives isolated in the baseline Downsian model – sometimes referred to as the *median voter theorem* – *do* operate in US politics, and are important.

Some examples:

- In US presidential races, candidates tend to advertise positions in the primary elections that are closer to the party median than the general electorate median. After winning their party's primary, there is a long history of US presidential candidates trying to "move towards the center," or to downplay the more left- or right- positions that they expressed in the primaries. This tendency is also evident in House and Senate races.[129]

[128] We could just as well term this the Hotelling-Downs model, or the Hotelling model, and some scholars do (Osborne, 1995). Hotelling first clearly stated the logic, even if he suggests the application to politics only in passing.

[129] A wonderful example: Immediately after winning the 2022 Republican Senate primary in New Hampshire, Don Bolduc pivoted to saying to that he had changed his mind and "come to the conclusion" that 2020 US presidential election was not, in fact, stolen. As he had "spent the last year claiming that it was," the *New York Times* described this as a "screeching U-turn."

References

Acemoglu, Daron. 2003. "Why not a political Coase theorem? Social conflict, commitment, and politics." *Journal of comparative economics* 31(4):620–652.

Acemoglu, Daron and James A Robinson. 2000. "Why did the West extend the franchise? Democracy, inequality, and growth in historical perspective." *The quarterly journal of economics* 115(4):1167–1199.

Acemoglu, Daron and James A Robinson. 2001. "A theory of political transitions." *American Economic Review* 91(4):938–963.

Acemoglu, Daron and James A Robinson. 2006. *Economic origins of dictatorship and democracy*. New York: Cambridge University Press.

Acharya, Avidit and Adam Meirowitz. 2017. "Sincere voting in large elections." *Games and Economic Behavior* 101:121–131.

Acharya, Avidit, Edoardo Grillo, Takuo Sugaya and Eray Turkel. 2020. "Electoral Campaigns as Dynamic Contests." Ms., Stanford University.

Acharya, Avidit, Eliot Lipnowski and João Ramos. 2021. "Optimal political career dynamics." Unpublished ms.

Achen, Christopher H. and Larry M. Bartels. 2017. *Democracy for realists*. Princeton, NJ: Princeton University Press.

Adler, Matthew D. 2019. *Measuring social welfare: An introduction*. Oxford University Press, USA.

Ahler, Douglas J and David E Broockman. 2018. "The delegate paradox: Why polarized politicians can represent citizens best." *The Journal of Politics* 80(4):1117–1133.

Alesina, Alberto. 1988. "Credibility and policy convergence in a two-party system with rational voters." *The American Economic Review* 78(4):796–805.

Alesina, Alberto and Enrico Spolaore. 2005. *The size of nations*. Mit Press.

Alesina, Alberto, Reza Baqir and William Easterly. 1999. "Public goods and ethnic divisions." *The Quarterly journal of economics* 114(4):1243–1284.

Ansolabehere, Stephen, Alan Gerber and Jim Snyder. 2002. "Equal votes, equal money: Court-ordered redistricting and public expenditures in the American states." *American Political Science Review* 96(4):767–777.

Ansolabehere, Stephen and Bernard L Fraga. 2016. "Do Americans Prefer Coethnic Representation? The Impact of Race on House Incumbent Evaluations." *Stanford Law Review* 68(6).

Ansolabehere, Stephen, James M Snyder Jr and Charles Stewart III. 2001. "Candidate positioning in US House elections." *American Journal of Political Science* pp. 136–159.

Arrow, Kenneth J. 1963. *Social choice and individual values*. New Haven, CT: Yale University Press.

Ashworth, Scott. 2012. "Electoral accountability: Recent theoretical and empirical work." *Annual Review of Political Science* 15:183–201.

Austen-Smith, David. 1984. "Two-party competition with many constituences." *Mathematical Social Sciences* 7(2):177–198.

Austen-Smith, David and Jeffrey Banks. 1988. "Elections, coalitions, and legislative outcomes." *The American Political Science Review* pp. 405–422.

Bafumi, Joseph and Michael C Herron. 2010. "Leapfrog representation and extremism: A study of American voters and their members in Congress." *American Political Science Review* pp. 519–542.

Bagnoli, Mark and Ted Bergstrom. 2005. "Log-concave probability and its applications." *Economic theory* 26(2):445–469.

Barber, Michael J. 2016. "Ideological donors, contribution limits, and the polarization of American legislatures." *The Journal of Politics* 78(1):296–310.

Baron, David P. 1994. "Electoral competition with informed and uniformed voters." *American Political Science Review* pp. 33–47.

Barro, Robert J. 1973. "The control of politicians: an economic model." *Public choice* 14:19–42.

Bartels, Larry M. 2018. "Partisanship in the Trump era." *The Journal of Politics* 80(4):1483–1494.

Bernhardt, Dan, John Duggan and Francesco Squintani. 2009. "The case for responsible parties." *American Political Science Review* 103(4):570–587.

Bernheim, B Douglas. 1984. "Rationalizable strategic behavior." *Econometrica: Journal of the Econometric Society* pp. 1007–1028.

Besley, Timothy. 2006. *Principled agents?: The political economy of good government*. Oxford: Oxford University Press.

Blackwell, David. 1965. "Discounted dynamic programming." *The Annals of Mathematical Statistics* 36(1):226–235.

Boehm, Christopher. 1999. *Hierarchy in the Forest: The Evolution of Egalitarian Behavior*. Cambrindge, MA: Harvard University Press.

Bonica, Adam. 2018. "DIME scores for congressional candidates for 1980-2018 election cycles." https://web.stanford.edu/ bonica/data.html.

Broockman, David E. 2016. "Approaches to studying policy representation." *Legislative Studies Quarterly* 41(1):181–215.

Brunell, Thomas. 2010. *Redistricting and representation: Why competitive elections are bad for America*. Routledge.

Callander, Steven. 2005. "Electoral competition in heterogeneous districts." *Journal of Political Economy* 113(5):1116–1145.

Calvert, Randall L. 1985. "Robustness of the multidimensional voting model: Candidate motivations, uncertainty, and convergence." *American Journal of Political Science* pp. 69–95.

Canes-Wrone, Brandice, David W Brady and John F Cogan. 2002. "Out of step, out of office: Electoral accountability and House members' voting." *American Political Science Review* 96(1):127–140.

Converse, Philip E. 2006. "The nature of belief systems in mass publics (1964)." *Critical review* 18(1-3):1–74.

Cox, Gary W. 1997. *Making votes count: strategic coordination in the world's electoral systems*. Cambridge University Press.

Cox, Gary W and Mathew D McCubbins. 2007. *Legislative leviathan: Party government in the House*. Cambridge University Press.

Dal Bó, Ernesto and Robert Powell. 2009. "A model of spoils politics." *American Journal of Political Science* 53(1):207–222.

Dixit, Avinash and John Londregan. 1995. "Redistributive politics and economic efficiency." *American political science Review* pp. 856–866.

Dixit, Avinash and John Londregan. 1996. "The determinants of success of special interests in redistributive politics." *the Journal of Politics* 58(4):1132–1155.

Downs, Anthony. 1957. *An economic theory of democracy*. New York: Harper.

Dube, Arindrajit, Oeindrila Dube and Omar García-Ponce. 2013. "Cross-border spillover: US gun laws and violence in Mexico." *American Political Science Review* 107(3):397–417.

Duclos, Jean-Yves, Joan Esteban and Debraj Ray. 2004. "Polarization: concepts, measurement, estimation." *Econometrica* 72(6):1737–1772.

Dutta, Bhaskar, Hans Peters and Arunava Sen. 2007. "Strategy-proof cardinal decision schemes." *Social Choice and Welfare* 28(1):163–179.

Duverger, Maurice. 1959. *Political parties, their organization and activity in the modern state*. Methuen.

Dwyer, Ryan J and Elizabeth W Dunn. 2022. "Wealth redistribution promotes happiness." *Proceedings of the National Academy of Sciences* 119(46).

Edgeworth, Francis Ysidro. 1897. "The pure theory of taxation, III." *The Economic Journal* 7(28):550–571.

Eguia, Jon X. 2013. "On the spatial representation of preference profiles." *Economic Theory* 52(1):103–128.

Elster, Jon. 1989. *Solomonic judgements: Studies in the limitation of rationality*. Cambridge University Press.

Esteban, Joan and Debraj Ray. 2011. "Linking conflict to inequality and polarization." *American Economic Review* 101(4):1345–74.

Esteban, Joan-Maria and Debraj Ray. 1994. "On the measurement of polarization." *Econometrica: Journal of the Econometric Society* pp. 819–851.

Fearon, James D. 1999. Electoral accountability and the control of politicians: selecting good types versus sanctioning poor performance. In *Democracy, Accountability, and Representation*, ed. Adam Przeworski, Susan C. Stokes and Bernard Manin. Cambridge University Press pp. 55–97.

Fearon, James D. 2011. "Self-enforcing democracy." *The Quarterly Journal of Economics* 126(4):1661–1708.

Fearon, James D, Macartan Humphreys and Jeremy M Weinstein. 2015. "How does development assistance affect collective action capacity? Results from a field experiment in post-conflict Liberia." *American Political Science Review* 109(3):450–469.

Ferejohn, John. 1986. "Incumbent performance and electoral control." *Public choice* 50(1):5–25.

Fiorina, Morris P. 2017. *Unstable Majorities: Polarization, Party Sorting, and Political Stalemate*. Hoover Press.

Fowler, Anthony, Seth Hill, Jeff Lewis, Chris Tausanovitch, Lynn Vavreck and Christopher Warshaw. 2021. "Moderates." Unpublished manuscript, August.

Francois, Patrick, Ilia Rainer and Francesco Trebbi. 2015. "How is power shared in Africa?" *Econometrica* 83(2):465–503.

Fudenberg, Drew and Eric Maskin. 1986. "The Folk Theorem in Repeated Games with Discounting or with Incomplete Information." *Econometrica* 54(3):533–554.

Fudenberg, Drew and Jean Tirole. 1991. *Game Theory*. Cambridge, MA: Mit Press.

Gambetta, Diego. 1993. *The Sicilian Mafia: The Business of Private Protection*. Cambridge, MA: Harvard University Press.

Gibbard, Allan. 1973. "Manipulation of voting schemes: a general result." *Econometrica* 41:587–601.

Gintis, Herbert. 2014. *The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences-Revised Edition*. Princeton University Press.

Gottlieb, Jessica. 2015. "The logic of party collusion in a democracy: Evidence from Mali." *World Politics* 67(1):1–36.

Groseclose, Tim and James M Snyder Jr. 1996. "Buying supermajorities." *American Political Science Review* 90(2):303–315.

Grossman, Gene M and Elhanan Helpman. 1992. "Protection for sale.".

Grossman, Gene M and Elhanan Helpman. 2001. *Special interest politics*. MIT press.

Hall, Andrew B. 2015. "What happens when extremists win primaries?" *American Political Science Review* 109(1):18–42.

Hall, Andrew B. 2019. *Who Wants to Run?: How the Devaluing of Political Office Drives Polarization*. University of Chicago Press.

Hall, Andrew B and Daniel M Thompson. 2018. "Who punishes extremist nominees? Candidate ideology and turning out the base in US elections." *American Political Science Review* 112(3):509–524.

Harsanyi, John C. 1967. "Games with incomplete information played by "Bayesian" players, I–III Part I. The basic model." *Management science* 14(3):159–182.

Hassan, Mai. 2024. "Coordinated dis-coordination." *American Political Science Review* 118(1):163–177.

Healy, Andrew and Neil Malhotra. 2013. "Retrospective voting reconsidered." *Annual Review of Political Science* 16:285–306.

Henderson, John. 2019. "Issue Distancing in Congressional Elections." ms., Yale University.

Henrich, Joseph, Maciej Chudek and Robert Boyd. 2015. "The Big Man Mechanism: how prestige fosters cooperation and creates prosocial leaders." *Philosophical Transactions of the Royal Society B: Biological Sciences* 370(1683):20150013.

Hersh, Eitan. 2020. *Politics is for power: How to move beyond political hobbyism, take action, and make real change*. Simon and Schuster.

Hill, Kim and A Magdalena Hurtado. 2009. "Cooperative breeding in South American hunter–gatherers." *Proceedings of the Royal Society B: Biological Sciences* 276(1674):3863–3870.

Hinich, Melvin J, John O Ledyard and Peter C Ordeshook. 1973. "A theory of electoral equilibrium: A spatial analysis based on the theory of games." *The Journal of Politics* 35(1):154–193.

Holmström, Bengt and Paul Milgrom. 1987. "Aggregation and linearity in the provision of intertemporal incentives." *Econometrica: Journal of the Econometric Society* pp. 303–328.

Holmström, Bengt and Paul Milgrom. 1991. "Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design." *Journal of Law, Economics and Organization* 7:24–52.

Hotelling, Harold. 1929. "Stability in competition." *The Economic Journal* 39(153):41–57.

Huber, John D and G Bingham Powell. 1994. "Congruence between citizens and policymakers in two visions of liberal democracy." *World Politics* 46(3):291–326.

Hylland, Aanund. 1980. "Strategy proofness of voting procedures with lotteries as outcomes and infinite sets of strategies." *Unpublished paper, University of Oslo.[341, 349]* .

Iyengar, Shanto, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra and Sean J Westwood. 2019. "The origins and consequences of affective polarization in the United States." *Annual review of political science* 22:129–146.

Kahneman, Daniel, Stewart Paul Slovic, Paul Slovic and Amos Tversky. 1982. *Judgment under uncertainty: Heuristics and biases*. Cambridge university press.

Kaneko, Mamoru and Kenjiro Nakamura. 1979. "The Nash social welfare function." *Econometrica: Journal of the Econometric Society* pp. 423–435.

Key, V.O. 1966. *The responsible electorate*. Cambridge, MA: Harvard University Press.

Kinder, Donald R and Nathan P Kalmoe. 2017. *Neither liberal nor conservative*. University of Chicago Press.

Krasa, Stefan and Mattias K Polborn. 2018. "Political competition in legislative elections." *American Political Science Review* 112(4):809–825.

Krehbiel, Keith. 1992. *Information and legislative organization*. University of Michigan Press.

Krehbiel, Keith. 2010. *Pivotal politics*. University of Chicago Press.

Kreps, David M. 1990. Corporate culture and economic theory. In *Perspectives on Positive Political Economy*, ed. James E. Alt and Kenneth A. Shepsle. Cambridge, MA: Cambridge University Press pp. 90–143.

Kujala, Jordan. 2020. "Donors, primary elections, and polarization in the united states." *American Journal of Political Science* 64(3):587–602.

Kuran, Timur. 1991. "Now out of never: The element of surprise in the East European revolution of 1989." *World Politics: A Quarterly Journal of International Relations* pp. 7–48.

Lalley, Steven P and E Glen Weyl. 2018. Quadratic voting: How mechanism design can radicalize democracy. In *AEA Papers and Proceedings*. Vol. 108 pp. 33–37.

Landemore, Hélène. 2020. *Open democracy*. Princeton University Press.

Ledyard, John O. 1984. "The pure theory of large two-candidate elections." *Public choice* 44(1):7–41.

Lenz, Gabriel S. 2013. *Follow the leader?: how voters respond to politicians' policies and performance*. University of Chicago Press.

Lindbeck, Assar and Jörgen W Weibull. 1987. "Balanced-budget redistribution as the outcome of political competition." *Public choice* 52(3):273–297.

Lizzeri, Alessandro and Nicola Persico. 2004. "Why did the elites extend the suffrage? Democracy and the scope of government, with an application to Britain's "Age of Reform"." *The Quarterly journal of economics* 119(2):707–765.

Luce, R Duncan and Howard Raiffa. 1957. *Games and decisions: Introduction and critical survey*. New York: Wiley.

Maslow, Abraham Harold. 1943. "A theory of human motivation." *Psychological review* 50(4):370.

Maynard Smith, John. 1982. *Evolution and the Theory of Games*. Cambridge university press.

McCarty, Nolan. 2019. *Polarization: What everyone needs to know®*. Oxford University Press.

McCarty, Nolan, Jonathan Rodden, Boris Shor, Chris Tausanovitch and Christopher Warshaw. 2019. "Geography, uncertainty, and polarization." *Political Science Research and Methods* 7(4):775–794.

McCarty, Nolan, Keith T Poole and Howard Rosenthal. 2009. "Does gerrymandering cause polarization?" *American Journal of Political Science* 53(3):666–680.

Meltzer, Allan H and Scott F Richard. 1981. "A rational theory of the size of government." *Journal of political Economy* 89(5):914–927.

Mirrlees, James A. 1971. "An exploration in the theory of optimum income taxation." *The review of economic studies* 38(2):175–208.

Morgenstern, Oskar and John Von Neumann. 1947. *Theory of games and economic behavior*. 2nd ed. Princeton, NJ: Princeton University Press.

Myerson, Roger. 2006. "Federalism and incentives for success in democracy." *Quarterly Journal of Political Science* 1:3–23.

Myerson, Roger B. 1995. "Analysis of democratic institutions: structure, conduct and performance." *Journal of Economic Perspectives* 9(1):77–89.

Myerson, Roger B. 2008. "The autocrat's credibility problem and foundations of the constitutional state." *American Political Science Review* 102(1):125–139.

Nakamura, Emi and Jón Steinsson. 2018. "Identification in macroeconomics." *Journal of Economic Perspectives* 32(3):59–86.

Nalebuff, Barry. 1986. "Brinkmanship and nuclear deterrence: the neutrality of escalation." *Conflict Management and Peace Science* 9(2):19–30.

Nash, John. 1953. "Two-person cooperative games." *Econometrica: Journal of the Econometric Society* pp. 128–140.

Nash, John F. 1950*a*. "The bargaining problem." *Econometrica* 18(2):155–162.

Nash, John F. 1950*b*. "Equilibrium points in n-person games." *Proceedings of the national academy of sciences* 36(1):48–49.

North, Douglass C and Barry R Weingast. 1989. "Constitutions and commitment: the evolution of institutions governing public choice in seventeenth-century England." *Journal of economic history* pp. 803–832.

North, Douglass C, John Joseph Wallis, Barry R Weingast et al. 2009. *Violence and social orders: A conceptual framework for interpreting recorded human history*. Cambridge University Press.

Oates, Wallace E. 1999. "An essay on fiscal federalism." *Journal of economic literature* 37(3):1120–1149.

Olson, Mancur. 1993. "Dictatorship, democracy, and development." *American political science review* 87(3):567–576.

Osborne, Martin J. 1995. "Spatial models of political competition under plurality rule: A survey of some explanations of the number of candidates and the positions they take." *Canadian Journal of Economics* 28(2):261–301.

Padró i Miquel, Gerard. 2007. "The control of politicians in divided societies: The politics of fear." *Review of Economic Studies* 74(4):1259–1274.

Palfrey, Thomas R. 1984. "Spatial equilibrium with entry." *The Review of Economic Studies* 51(1):139–156.

Pearce, David G. 1984. "Rationalizable strategic behavior and the problem of perfection." *Econometrica: Journal of the Econometric Society* pp. 1029–1050.

Piketty, Thomas and Emmanuel Saez. 2013. Optimal labor income taxation. In *Handbook of public economics*. Vol. 5 Elsevier pp. 391–474.

Polborn, Mattias K and James M Snyder. 2017. "Party polarization in legislatures with office-motivated candidates." *The Quarterly Journal of Economics* 132(3):1509–1550.

Poole, Keith T. 2005. *Spatial models of parliamentary voting*. Cambridge University Press.

Poole, Keith T and Howard Rosenthal. 1985. "A spatial model for legislative roll call analysis." *American journal of political science* pp. 357–384.

Poole, Keith T and Howard Rosenthal. 1991. "Patterns of congressional voting." *American journal of political science* pp. 228–278.

Poole, Keith T and Howard Rosenthal. 2000. *Congress: A political-economic history of roll call voting*. Oxford University Press on Demand.

Poole, Keith T and Howard Rosenthal. 2001. "D-nominate after 10 years: A comparative update to congress: A political-economic history of roll-call voting." *Legislative Studies Quarterly* pp. 5–29.

Poole, Keith T and Howard Rosenthal. 2017. *Ideology & congress: A political economic history of roll call voting*. Routledge.

Powell, G Bingham. 2000. *Elections as instruments of democracy: Majoritarian and proportional visions*. New Haven, CT: Yale University Press.

Powell, Robert. 1990. *Nuclear deterrence theory: The search for credibility*. Cambridge University Press.

Powell, Robert. 2007. "Defending against terrorist attacks with limited resources." *American Political Science Review* 101(3):527–541.

Powell, Robert. 2009. "Sequential, nonzero-sum "Blotto": Allocating defensive resources prior to attack." *Games and Economic Behavior* 67(2):611–615.

Rawls, John. 1971. *A Theory of Justice*. Harvard University Press.

Rehfeld, Andrew. 2005. *The concept of constituency: Political representation, democratic legitimacy, and institutional design*. Cambridge University Press.

Robbins, Lionel. 1938. "Interpersonal comparisons of utility: a comment." *The Economic Journal* 48(192):635–41.

Roberts, Kevin WS. 1980. "Interpersonal comparability and social choice theory." *The Review of Economic Studies* pp. 421–439.

Rodden, Jonathan A. 2006. *Hamilton's paradox: the promise and peril of fiscal federalism*. Cambridge University Press.

Rodden, Jonathan A. 2019. *Why Cities Lose: The Deep Roots of the Urban-Rural Political Divide*. New York: Basic Books.

Roemer, John E. 1994. "A theory of policy differentiation in single issue electoral politics." *Social Choice and Welfare* 11(4):355–380.

Roemer, John E. 1996. *Theories of distributive justice*. Harvard University Press.

Roemer, John E. 1997. "Political–economic equilibrium when parties represent constituents: the unidimensional case." *Social Choice and Welfare* 14(4):479–502.

Roemer, John E. 2009. *Political competition: Theory and applications*. Cambridge, MA: Harvard University Press.

Roessler, Philip. 2016. *Ethnic politics and state power in Africa: The logic of the coup-civil war trap*. Cambridge University Press.

Sánchez de la Sierra, Raúl. 2020. "On the origins of the state: Stationary bandits and taxation in Eastern Congo." *Journal of Political Economy* 128(1):32–74.

Satterthwaite, Mark A. 1975. "Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions." *Journal of Economic Theory* 10(2):187–217.

Schelling, Thomas C. 1960. *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.

Schelling, Thomas C. 1978. *Micromotives and macrobehavior*. New York: WW Norton & Company.

Schelling, Thomas C. 1996. *Arms and Influence*. Yale University Press.

Sen, Amartya. 1970. *Collective choice and social welfare*. Harvard University Press.

Sen, Amartya. 2017. *Collective choice and social welfare: An expanded edition*. Cambridge, MA: Harvard University Press.

Shotts, Kenneth W. 2001. "The effect of majority-minority mandates on partisan gerrymandering." *American Journal of Political Science* pp. 120–135.

Smithies, Arthur. 1941. "Optimum location in spatial competition." *Journal of Political Economy* 49(3):423–439.

Snyder, James M. 1994. "Safe seats, marginal seats, and party platforms: The logic of platform differentiation." *Economics & Politics* 6(3):201–213.

Snyder, James M., Michael M. Ting and Stephen Ansolabehere. 2005. "Legislative bargaining under weighted voting." *American Economic Review* 95(4):981–1004.

Snyder, James M. and Stephen Ansolabehere. 2008. *The End of Inequality: One Person, One Vote and the Reshaping of American Politics*. WW Norton & Company.

Stephanopoulos, Nicholas O. 2012. "Spatial Diversity." *Harvard Law Review* pp. 1903–2010.

Stone, Peter. 2011. *The luck of the draw: The role of lotteries in decision making*. Oxford University Press.

Tabellini, Guido. 2008. "The scope of cooperation: Values and incentives." *The Quarterly Journal of Economics* 123(3):905–950.

Tausanovitch, Chris and Christopher Warshaw. 2013. "Measuring constituent policy preferences in congress, state legislatures, and cities." *The Journal of Politics* 75(2):330–342.

Tiebout, Charles M. 1956. "A pure theory of local expenditures." *Journal of political economy* 64(5):416–424.

von Neumann, John and Oskar Morgenstern. 1947. *Theory of games and economic behavior (2nd rev. ed.)*. 2nd ed. Princeton University Press.

Weingast, Barry R. 1979. "A rational choice perspective on congressional norms." *American Journal of Political Science* pp. 245–262.

Weingast, Barry R. 1994. "Reflections on distributive politics and universalism." *Political Research Quarterly* 47(2):319–327.

Weingast, Barry R. 1997. "The political foundations of democracy and the rule of law." *American Political Science Review* 91(2):245–263.

Weingast, Barry R and William J Marshall. 1988. "The industrial organization of Congress; or, why legislatures, like firms, are not organized as markets." *Journal of Political Economy* 96(1):132–163.

Wittman, Donald. 1983. "Candidate motivation: A synthesis of alternative theories." *The American political science review* pp. 142–157.

Wittman, Donald. 1990. Spatial strategies when candidates have policy preferences. In *Advances in the Spatial Theory of Voting*, ed. James M. Enelow and Melvin J. Hinich. Cambridge University Press chapter 4, pp. 66–98.

Wittman, Donald A. 1973. "Parties as utility maximizers." *The American Political Science Review* 67(2):490–498.