JOURNAL OF THEORETICAL POLITICS

Original Research Article

Social Power and Non-cooperative Game Theory

Journal of Theoretical Politics
1–18

DOI: 10.1177/09516298221081810
journals.sagepub.com/home/jtp
SAGE

William Bosworth
Philosophy and Economics, SPIR, Australian National University

Abstract

This paper defends the use of non-cooperative game theory for analysing questions of governance. To do so it posits a way of extending the resource account of social power from cooperative games to noncooperative games in a way that side steps a range of criticism. This involves identifying tipping points in the reputations of certain agents for paying and punishing those in their thrall. These tipping points are what give threats and offers their credibility in the absence of enforcement mechanisms and stabilise the distribution of social resources in society.

Keywords

Power, governance, bargaining, bayesian games, reputation, tipping points

The question of who governs is arguably *the* question of political science. It was posed by Robert Dahl (2005) [1961] to empirically assess whether inequalities in the market translated into inequalities in governance at City Hall. If true, Dahl argued Marx was essentially right. The formal equality bestowed to citizens through the vote was little more than an ideological smokescreen. If not, there was hope yet for American capitalism.

Dahl's study ended with optimism. He used observable measures to identify policy conflicts amongst his city's (New Haven, Connecticut) representatives and counted who was ultimately successful and who was not. It was observed that different groups were decisive across the discrete policy areas of urban redevelopment, public education, and political nominations. Dahl concluded New Haven was a 'polyarchy', where the

Corresponding author:
Dr William Bosworth, Philosophy and Economics, SPIR, Australian National University, Canberra, ACT 0200, Australia.
Email: william.bosworth@anu.edu.au

many (poly) govern (arkein) and implied radical institutional change should be accordingly tempered.

The rational choice interpretation, however, suggests genuine political conflict may never reach city hall to begin with. It takes the mobilisation of resources to secure representation and individual rationality regularly dictates sub-optimal outcomes for groups. The very difference between the inequalities of the market and the inequalities of governance, by this interpretation, boils down to questions of coordination. Consumers may collectively have the aggregate resources to secure representatives to advocate tighter regulations on business, for instance, but because of incentives for freeriding that come with such large groups, cannot mobilise those resources to do so. A similar kind of explanation can be proffered for why the poorest 99.9% do not expropriate the wealthiest 0.1%, how 18th Century slaveholders in the Caribbean managed to maintain control over their brutal sugar plantations despite being outnumbered 10 to 1 (Petley, 2018), and why some women consciously perpetuated oppressive patriarchal conventions they knew hindered their and their daughters' life prospects (Wollstonecraft, 1993: 276). Coordination problems like this are described with non-cooperative game theory.

Yet there is a theoretical loose end that has left the rational choice approach vulnerable to criticism on grounds ranging from ideological bias (Wolfinger, 1971: 1078; Polsby, 1980: 96-7; Connolly, 1974: 126-30; Barry, 2002; Lukes, 2005: e.g. 110-11) through to incommensurability (Morriss, 2002: 138-44). Pluralists question why the inequalities in the ability to coordinate wealth should matter in light of high levels of voter coordination. 'Elitists' on the other hand argue the interpretation focuses too heavily on the coordination problems of the oppressed, glossing over the culpability of elite oppressors like capitalists, colonialists, and men. It begs the question which kinds of strategic non-action are relevant to the question of governance and how they might exacerbate or mitigate other coordination problems.

This paper proposes an answer by joining the rational choice analysis of the cooperative games initially used to describe governance with the non-cooperative games characterising coordination problems more generally. This is complicated by the fact that cooperative games assume over the possibility of sub-optimal equilibrium with the axiom of Pareto optimality. Without further elaboration, however, different strategies for measuring governance will be underdetermined, opening the door to charges of partiality (e.g. see Wolfinger 1971: 1078). Radicals will focus on the coordination problems associated with wealth and clientalism; conservatives will focus on voter coordination. Without further refinement here, empirical studies risk generating answers to who governs that are tinged with ideology.

The cooperative game

John Harsanyi (1962a,b) based the original rational choice interpretation of governance on cooperative bargaining games. His formal analysis suggests the opportunity costs associated with getting others to do what they would *prefer not to* are what make the question of governance in and of itself important. The ability to get others to do what they would rather not is roughly Weber's (1978: 53) definition of social power as the ability to overcome the resistance of others (see also Barry, 1989a: 272; 2002: esp. 161). Agent A has social power over B to the extent they can get B to do X where X is an action B assigns disutility to. This is a subset of Dahl's (1957) own interpretation, but Harsanyi's rational choice analysis suggests it is at least sufficient (if not necessary) for making who governs the important question that it is.[1]

The means to get B to do X despite preferring not to are credible threats of punishment in the event B does not do X and credible offers of rewards in the event they do. This can range from electoral sanction in the way of votes and campaign contributions, to the punishment of citizens by way of incarceration for breaking the law. Harsanyi adapts Nash; (1953) cooperative bargaining game to bring the strategic dynamics associated with both threats and offers under a unified analytic framework. The premise is that certain kinds of agents A are locked into tacit bargaining games with other kinds of agents B over the mixed strategy *likelihood* of B performing X, where A wants B to perform X but B themselves do not. The agents assign utilities to the various likelihoods and also to the point where the tacit bargaining process breaks down. The idea is that credible threats and offers will manipulate the utility assigned to the disagreement points in the cooperative game and thereby influence what is rational to tacitly agree to when it comes to the bargain over likelihoods.

The amount of A's power over B with respect to X is A's ability to alter this likelihood with a threat or offer. That is, the amount of power is p2-p1 where p1 is the initial probability of B doing action X and p2 the probability after A has rationally exercised any threats and offers available to them. The value of p_2 is defined as

$$p_2 = p_1 + \frac{\mathrm{r}+\mathrm{t}}{2x} + \frac{r* \ -t*}{2x*}$$

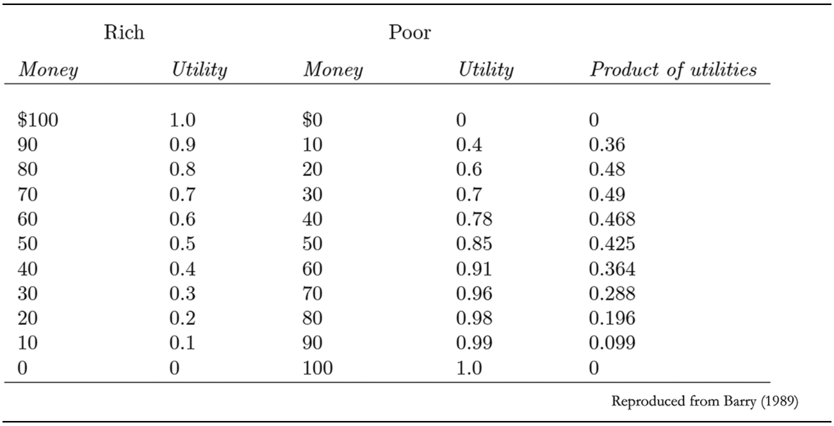
where x is the disutility B associates with doing X, x* the utility A associates with B doing X, r the value of rewards B would receive from A if A were rational, r* the costs of those rewards to A, t the disutility B associates with A's rational punishment, and t* the cost of it to A. Harsanyi (1962b) extends the idea in a complementary paper to the *n*-person case.

The account is justified by the Nash solution to the bargaining game. For the two-player case, A is assigned a payoff *x* and B a payoff *x** for every likelihood. There is also a payoff assigned to the event of disagreement (*d,d**). Nash proved there is a unique solution to the bargaining game, so defined, satisfying the relatively thin axioms of Pareto optimality, individual optimality, independence of irrelevant alternatives, symmetry, and scale covariance (Nash, 1953). The solution is the point that maximizes

$$(\mathrm{x} - \mathrm{d})(\mathrm{x}^{*} - \mathrm{d}^{*}).$$

For Harsanyi's interpretation of power the probability agent B performs X is the object of the bargain. The disagreement point is set by threats and offers (where an offer increases the disagreement in the sense it can be interpreted as a threat *not* to reward).

Nash's bargaining solution shows here the normative importance of the *relative* costs of rewards and punishments. A rich billionaire threatening legal action against a struggling tenant, for example, is in a considerable bargaining advantage because the tenant has greater opportunity costs associated with litigation given their meager resources. When splitting $100 valuing it *less* (because the agent is less desperate) is a considerable bargaining advantage,



Rich		Poor		
Money	*Utility*	*Money*	*Utility*	*Product of utilities*
$100	1.0	$0	0	0
90	0.9	10	0.4	0.36
80	0.8	20	0.6	0.48
70	0.7	30	0.7	0.49
60	0.6	40	0.78	0.468
50	0.5	50	0.85	0.425
40	0.4	60	0.91	0.364
30	0.3	70	0.96	0.288
20	0.2	80	0.98	0.196
10	0.1	90	0.99	0.099
0	0	100	1.0	0

Reproduced from Barry (1989)

The largest product is (0.7)(0.7) = .49 where Poor gets $30 and Rich gets $70. This is only the bargaining solution, however, when the disagreement points d and d* are fixed at 0. If the Rich can threaten to punish the Poor in a way that makes Poor's disagreement point disproportionately worse, it will often be rational for Rich to do so even though the punishment is costly. Say $d = -0.05$ and $d^* = -0.5$. The Nash Product for the $70/30 split would be $(0.7 + 0.05)(0.7 + 0.5) = 0.9$, but the product for the $80/20 split $(0.8 + 0.05)(0.6 + 0.5) = 0.935$ is now higher. The same goes for bargaining over the likelihood of B performing X. It will usually pay A to decrease the disagreement point of B via a threat or offer when it decreases B's disagreement point more than A's. The old adage, "This will hurt you more than me" is valid in this context (see also Rubinstein, 1982).

This is why who governs is a distinctly important question. Those with relatively more social power will have lower opportunity costs associated with discrete threats and offers. It will therefore be sub-optimal for those with relatively less social power to be proactive in making threats and offers to the socially powerful given the opportunity costs for honouring them will be disproportionately higher. It is also rational for the socially powerful to exercise threats *rather* than offers. If the punishment associated with a threat will deal out higher costs to B than the costs A associates with performing the punishment, it is rational for A to increase the severity of the threat up to the point where $t^*(1 - p_2) + r^*(p_2) = x_2(p_2)$. This is because the threat will *decrease* the likelihood the costly punishment will be required by *increasing* the likelihood B will comply. The same is not true for offers given the higher the offer, the more likely the offer will be required to reward B's (likelier) compliance. The bargaining analysis therefore suggests those with relatively more social power will be incentivized to use threats against those who do not and those with relatively less social power will be incentivized not to. The identity of who governs is therefore important: if it turned out a diverse range of large groups can coordinate to pass coercive law, then the bargaining disadvantage associated with the inequalities in wealth may be offset. The disadvantages are otherwise hard to justify and are, if anything, exacerbated.

Resources

Harsanyi suggests social power can be measured indirectly by counting the aggregate number of resources (like money) groups can use to manipulate the disagreement points of the bargaining game. The rational choice approach, expanded to non-cooperative games, suggests that social power of groups should be measured by the raw aggregate of the group members' resources, as a proxy, and then qualified by any coordination problems the groups face (see Dowding, 1996; 2019). In theory, if the poorest 99.9% could pool and then mobilise their collective wealth, they could reverse the myopic bargaining inequalities of wealth favouring the top 0.1% without recourse to the vote and legislation. But this is unlikely due to the crippling coordination problems associated with such large groups (Olson, 1965). To answer the question posed in the introduction, then, the coordination problems relevant to the question of governance are the sub-optimal equilibria associated with the mobilisation of these resources.

The analysis cannot end here though since what constitutes a resource in the appropriate sense is unclear. Adding reputation as a resource, for example, has been called a "fudge factor" and the treatment of resources, generally, as epicycling characteristic of degenerative scientific paradigms (Barry, 2003: 325). The idea is that a resource is whatever can be used to cover the costs of increasing or decreasing an agent's expected utility. Peter Morriss (2002: 139) argues that resources in this sense are not good proxies for the measurement of social power because "studying resources is every bit as complicated, and indirect, as studying power itself."

It is plausibly even more complicated for social power. Harsanyi (1968a: 71), for one, thought resources like affection, legitimate authority, and information were also important to measure in addition to the resources that cover the costs of rewards and punishments. Yet it is unclear how we should measure legitimacy and information relative to one another, let alone how we would measure them relative to resources like wealth, votes, and basic legal rights. A benevolent master's great affection for their slave does not mitigate the normatively problematic bargaining asymmetries between the two (Pettit, 1997). The slave may be relatively lucky because of their master's nature, but they do not govern in the normatively relevant sense: they have nothing like a threat of legal sanction to bargain with.

Recall that the significance of who governs is captured by the bargaining game over the likelihood of B doing something they would *prefer not* to. Only resources for manipulating the game's disagreement point are relevant to this end. Insofar as a provision of information gets B to do Y by making them *prefer* doing Y, it is not in itself relevant in the way a threat that gets B to perform X despite B *still preferring not* to do X is. Doctors influence patients *to prefer* taking medicine with the information that the medicine will be beneficial to their health. This kind of influence is about changing what people prefer rather than getting them to do X despite preferring not to. It will not lead to the normatively pertinent bargaining inequalities that motivate the question of governance, nor will it mitigate them. In the same way, it is difficult to see how inequalities in legitimate authority – like the authority of a scientific expert – are normatively problematic unless they are won or sustained by coercive threats or bribes (see Barry, 2002: 161; Arendt, 1954: 93).

Even if we restrict the resources we are interested in to only those required to cover the costs of punishments and rewards, questions are nevertheless still begged without further elaboration. Such resources are not clear empirical datum in the way a "chain of office or palace" are because they count as "resources only if others recognize them as such: if the things that can be used to provide are valued [or feared] by the potential recipients" (Morriss, 2002: 139). The vote is only a decisive resource if there are policymakers who value being rewarded and fear being punished by voters for their actions in office. If policymakers are more concerned with pleasing financial interests who can fund their re-election campaigns, or even their life after politics, then the vote will be considerably less decisive than wealth. This leads Morriss (2002: 144) to the view that the debate between the pluralists and elitists in political science is entirely misconstrued in the absence of "a theory of the political process that allows us to evaluate these divergent sorts of resources".

Information, legitimate authority, and affection may all conspire to influence policymakers to either value or disregard the vote. Some argue that what workers value for rewards and punishments (e.g. capitalist wealth), and their non-action to do more to find what really is in their interest, is itself a product of social power (Lukes, 2005: 144). Narrowing the question of governance to who can cover the costs of rewards and punishment, then, may not sufficiently abstract away from other influences on the expected utility of agents. Measuring social power and governance in terms of resources, in other words, would appear to arbitrarily ignore certain non-actions. If true, our measurement of social power would be normatively partial. While some suggest this is inevitable (e.g. Lukes, 2005: 111; Connolly, 1988), the remainder of this article looks to develop a theory for a non-arbitrary abstraction.

Credibility

The "natural" contribution of rational choice to the analysis of social power is to capture the interrelated dynamics of threats and offers (Barry, 1989: 226). These dynamics are what motivate the importance of who governs. Yet the resources that are used to cover the costs of the rewards and punishments appear to beg the same problems of bias and incommensurability the rational choice approach was designed to overcome. In this section I suggest the move from cooperative to non-cooperative games opens up the analytic possibility of overcoming these problems.

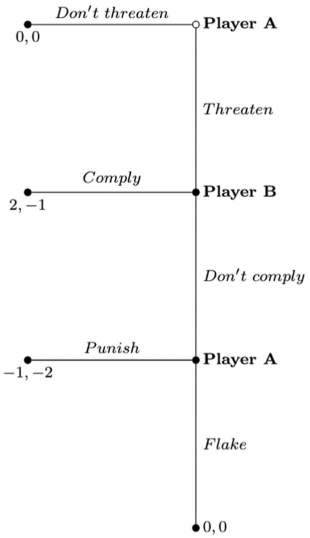
The key to defending the approach here is to factor an aspect of the criticism of the resource account into the account itself. To repeat, the resources used to cover the cost of rewards and punishments depend on being feared or "valued by potential recipients". Individuals will collect resources instead of consume them because they are feared or valued by potential future recipients. This implies potential recipients will have good reason to try and take an individual's resources. Having a resource therefore implies there is an effective threat or offer already in place to decisively exclude others from consuming (if they value) or destroying (if they fear) it. A government needs to fear the consequences of wide scale election fraud for the vote to count as a genuine resource; policymakers need to fear retaliation if wealth is not simply to be confiscated when offered as a bribe; there needs to be widespread fear of punishment for trespass if land is to be said to belong to anybody; governments need to fear punishment for violating basic rights like free speech for the right to free speech to count as a resource; and so on. The primitive concept, in other words, must be the credible threat or credible offer.

We measure resources as a proxy for an agent's ability to issue threats and offers. But social resources will themselves be the product of threats and offers. We therefore require a theory that can explain how threats and offers can be stably distributed in a credible way. The cooperative games that characterize Nash's threat game simply assume all threats and offers will be credible and do not interrogate the strategic reasoning that renders them credible in the first instance. As Nash (1953: 130) put it,

> "Supposing A and B to be rational beings, it is essential for the success of the threat that A be compelled to carry out his threat T if B fails to comply. Otherwise it will have little meaning. For, in general, to execute the threat will not be something A would want to do, just of itself… we must assume there is an adequate mechanism for forcing the players to stick to their threats and demands once made; and one to enforce the bargain, once agreed."

Every agent has the ability to issue an infinite number and range of threats and offers. But our abilities to issue *credible* threats and offers are subject to this assumed mechanism. It is this commitment mechanism that therefore determines who has social power and who does not. I could threaten to nuke the whole world if I did not receive my morning cup of coffee – anybody could – but it would not be credible. First, the punishment would be so costly given the punisher would also perish. Second, the number and range of threats and offers (to military chiefs and the like) required to carry it out would be equally incredible.

The Nash solution concept for non-cooperative games, however, does not in itself explain the way in which certain agents will have *more* of an ability to issue credible threats and offers than other agents. While the Nash equilibrium is a necessary analytical device for explaining the relevant commitment mechanism, it is not in itself sufficient. For example, if we assume that punishment is costly, the following extensive form game appears to capture the non-cooperative dynamics of the threat game,



While B complying is a Nash equilibrium when A's strategy is to threaten and punish, it is not sub-game perfect, which is to say A punishing is not a Nash equilibrium if the game were to begin at node 3. Why should A administer a costly punishment when the reason they raised the prospect of punishment in the first place has not been satisfied? We should flake at node 3 given we presumably assign some utility to our resources and usually assume there is little expressive value to paying or punishing in and of itself. The straight line down is the path of sub-game perfection (i.e. it is a Nash equilibrium at every node if we assume that node is a game unto itself). While threats and offers are asymmetric in the sense that it is costly to honour non-compliance with a threat whereas it is not for an offer, the offer game still presents the same puzzle. That is, it would not be sub-game perfect for A to honour a costly offer, given A will have already got what it wanted out of B if B has complied.

There are of course a range of commitment mechanisms in society for getting around this. For one, an individual could take out a large bet that they will honour their threat in the event of non-compliance. They thereby stand to make a large loss if they do not honour the threat, thereby making it credible. More common is the use of legal contracts to likewise commit to payment through the threat of legal punishment in the event the contract is not honoured. These solutions, however, beg the question how we can rely on the credible threats of law enforcement and credible offers of betting agencies in the first place. What makes a state's domestic threats credible?

The most general credibility inducing mechanism is reputation. It is primitive in the sense that it explains all other credibility-inducing mechanisms. It requires no preconditions like betting agencies or a dynamic legal system that will be contingent on the society's pre-existing power structure. We stake our reputation on our threats and offers. If we flake on our commitments, and this is observed, then onlookers will update their beliefs in a way that makes it common knowledge we are not of a type that prefers to punish (for the sake of a reputation or whatever) in node 2 of the game. Any future threat is no longer credible. If you do not honour your threats/offers, you compromise your future ability to issue new credible threats/offers.

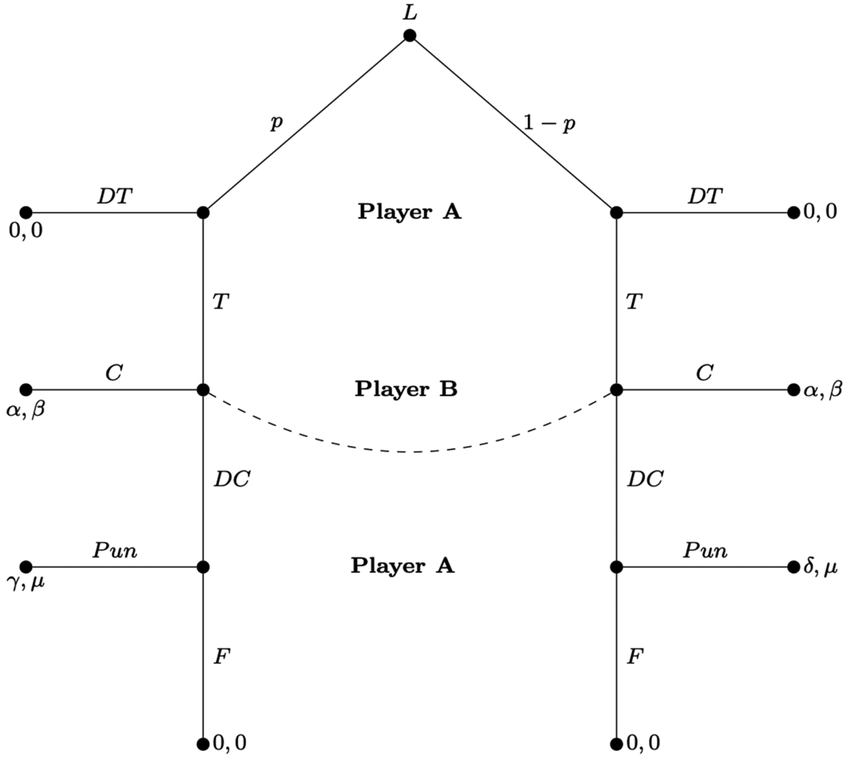
If the extensive form game above is repeated a billion times, it seems clear that building a reputation is an optimal strategy. If we preserve the assumption of perfect information, however, this is simply not the case (Selten, 1978). Via backward induction, we analyse the last game in the series (i.e. the billionth iteration). This game has the same structure as the extensive form threat game already discussed, with a subgame perfect equilibrium of Agent A flaking on their threat. So the outcome is already settled – when we get to the last game, now, it is essentially a dead rubber. We can therefore infer that this last game should be ignored in assessing the utility of building a reputation. But then the second last game (the billion-1th) becomes, in effect, the new last game. It will again have a subgame perfect equilibrium of A flaking on their threat. It too will become a dead rubber – and so on and so forth all the way through the billion iterations to the first game. With perfect certainty there is little incentive to build reputations in finite games by doing anything other than flaking on a threat.

It should be flagged that if the game is repeated an *infinite* number of times the folk theorem does suggest any finite sequence of actions (including rewards and punishment) will be part of a sub-game perfect equilibrium (e.g. Fedenberg and Maskin, 1986). Given the sheer number of equilibria for infinitely repeated games, however, the question is begged why commitment mechanisms like legal contracts are stable in the sense agents can predictively rely on them to honour successful offers and failed threats. The folk theorem suggests we can rationalize flaking in the next iteration *and also* rationalize not flaking; we can rationalize *making* threats and offers and at the same time rationalize not making them *for any agent*. The folk theorem, then, cannot explain the mechanism that leads to certain agents having the ability to issue *more* credible threats and offers than others.[2]

Wood for the crucifixion

The only option, it seems, is to relax the assumption of certainty that characterised the non-cooperative games considered in the previous section. It was not until after the formal interpretation of social power was first posed that it was shown possible to relax the assumption of complete certainty in a plausible way. This analysis was introduced by Harsanyi (1967) himself with the notion of the Bayesian Nash equilibrium, but Harsanyi never returned to his theory of social power to elaborate on its implications. There is perhaps good reason the two arguments have not been hitherto joined. In this section I will show how solution concepts for Bayesian games gives us a way to think of a dynamically updating environment in light of prior decisions that is amenable to the analysis of credible threats. At the same time, however, it reveals a problem for squaring it with the cooperative games used to describe the relations of governance and social power.

Assume A and B could be different types of player in the sense that they could have different payoffs. Say B does not know which type A is *but both A and B have common priors concerning the likelihood* that they are of particular types. This is modelled as a node L in an extensive form game that is described as a 'natural lottery',



$\alpha > 0 > \beta > \mu, \quad \delta > 0, \quad \gamma < 0$

It is common knowledge that player A is a type (type 1) that actively gains from punishment with probability 1 – p.[3] That is to say, player A gets δ utility from punishing with probability 1 – p and γ with probability p. While the above game models one-sided uncertainty on the part of B, the same can be straightforwardly done for A. It could also be done for cases where B is uncertain of the cost of flaking for A. The inequalities in this instance look like $\alpha > 0 > \beta > \mu$, $\gamma < 0$, and $\delta > 0$.

It has been shown that with an elaboration of the Nash equilibrium for sequential games, the slightest uncertainty on the part of B that A could be type 1 (1 – p >0) is enough to break the backward induction result described above (Kreps and Wilson, 1982). Assuming B reacts to A's punishment by updating their belief with Bayes' theorem, even where the likelihood is close to 0, and B thinks that it is highly likely A will only punish for reputation's sake, threatening is still credible in an iterated version of the threat game given a reputation for building a reputation for punishment is still a reputation for punishment (given A will still need to punish to build it) (Kreps and Wilson, 1982).

Kreps and Wilson's account of sequential equilibrium here captures an important aspect of reputation and credibility that is necessary for unpacking the concept of a credible threat and credible offer. But it is still not enough to complete the resource account of social power. According to the sequential equilibrium analysis, if A flakes, then B must update their belief such that the probability of A being type 1 is 0. It is the possibility of updating our beliefs with certainty that gives the game its sequential equilibrium (Myerson, 1991: 168-77). **If there is even the *slightest* probability A is type 1 after a default, there is no incentive for A to punish in any game and no incentive therefore for B to comply.** As Kreps and Wilson (1982: 262) put it, "the remarkable fact about this equilibrium is that even for very small *p*, the "reputation" effect soon predominates".

But this is not how we process reputation-destroying acts when considerations of resources are in play. Player A might still have the behavioural tendencies of type 1 but not have the resources to punish. A king might actively enjoy punishing their prisoners by crucifixion but be temporarily out of wood. To update p = 1 given the king's failure could lead one into doom once the monarch's next shipment of logs arrives. A's strategy may well be type 2 'Always punish if possible', but B will not be able to verify this without an independent theory about A's resources.

In Bayesian game theory the only way we can represent uncertainty over resources is in the utility function of the types (Harsanyi, 1967). So if a certain type knows they are out of wood, the strategy of punish by way of crucifixion needs to be represented as infeasible by some sufficiently low payoff like -10^{1000} such that it is never reasonable for the type to play it.

If resource levels are temporary, as in the case of the king, this type cannot stick to the player. Yet representations of payoffs in terms of utility *do* stick with sequential updating of belief. It is an error to think being temporarily out of wood is evidence the king will not go back to crucifying their prisoners once a new shipment arrives. It is therefore a mistake to ever update one's belief that p = 1 given the uncertainty associated with resources. Without this move the reputation effect in Bayesian games disappears.

Momentary thresholds and tipping points

So an agent may actively gain from punishment but have insufficient resources to punish. The agent may actively want to honour their offers independent of considerations of reputation but not have the resources to do so. This requires further caveats given A does not reveal there is 0 probability they actively benefit from punishing independent of reputation-effects because they may simply be out of resources.

My claim is that all that is required for the reputation effect is the presence of a certain group who are already predisposed to help A pay/punish others by virtue of threats/offers (be they explicit or implicit) previously issued by A. This group is already in A's thrall. A has already made its threats and offers to this group. It is the active and measurable exercise of social power that stabilises an individual's resources that the criticisms of the resource approach to social power missed. This group constitutes the sum total of actions individuals wouldn't otherwise have done to help A honour their threat/offer with another member because of the threat/offer A has issued (either explicitly or tacitly) to them.

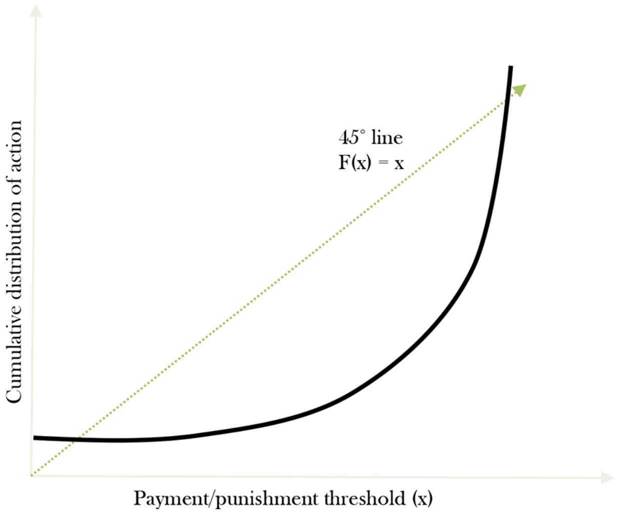
Now say this group observes A's possibly temporary default on a threat/offer. Within this class there will be individuals **whose threats/offers are equal or greater than the cost to honour than the punishment/reward just defaulted on**. This means A's threat/ offer over them is *momentarily* incredible, so A cannot rely on them in that moment.

It is rational to believe that *in the moment* p = 1 for this sub-group of individuals whose punishment/payment costs *more than* the punishment/payment just defaulted on. While momentary, this belief will often be enough to justify a run on A's credibility in the minds of the onlookers. It is the same sort of phenomenon that occurs in the minds of individuals during bank-runs and revolutions.

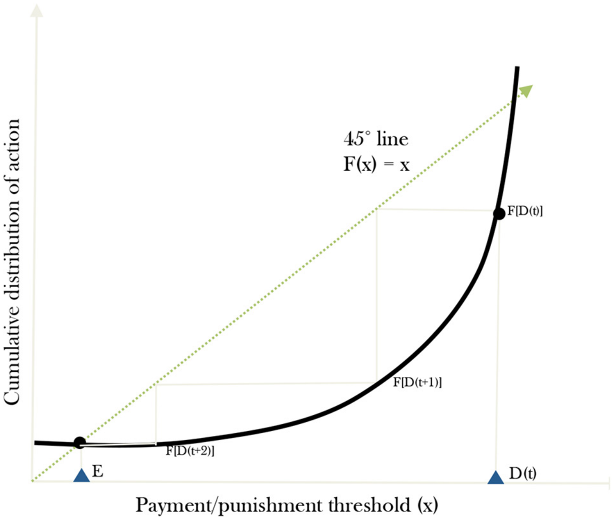
A mafia don's muscle is an important resource for the mafia to punish non-compliance. If the don's muscle observes the don flake on a threat or offer, they are likewise justified in believing she will temporarily flake on the threats/offers she made to them if they are greater or equal in cost This will exacerbate the don's precarious situation and could lead to a run on their credibility where revolt is *at this moment* considered tenable by rivals by virtue of new doubts about the credibility of the threats to the don's capos.

That is to say, certain defaults will be *tipping points* (see Schelling, 1978) for an individual's resources and the desire to avoid these tipping points will lead to credibility-inducing mechanisms in the threat/offer game. This is a momentary tipping that will not be based on observed sequential moves.

In fact, the ability to pay/punish is crucially not taken in terms of opportunity-cost or disutility to the payer/punisher at all, but rather the raw number of actions A has under their thrall to punish/pay for (non)compliance. We can capture what is going on with a simple threshold graph. The x-axis represents the number of actions required to punish or pay each threat or offer. The y-axis represents the cumulative number of actions A has under their thrall.



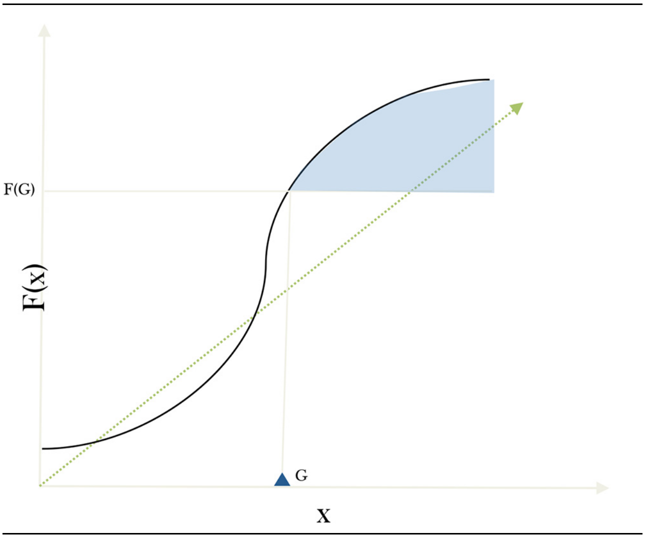
This graph is a hypothetical theoretical (and abstract) representation of A’s social resources. The 45-degree line represents when an individual’s own compliance is decisive for covering the cost of her own punishment or payment (when F(x) = x). Given they are decisive it does not make sense for them to help punish or reward themselves, therefore they will withdraw their compliance. We can use this line to find the equilibrium E for when onlookers observe the default D(t),


Cumulative distribution of action
45° line
F(x) = x
F[D(t)]
F[D(t+1)]
F[D(t+2)]
E
D(t)
Payment/punishment threshold (x)

F[D(t)]–F[E] will be large when an agent has to start from close to the origin point and work up underneath the 45-degree line (F(x) = x). A well-publicized default on paying for compliance on D(t) will be catastrophic for the credibility of a large chunk of their threats and offers. The potential for tipping points of credibility is what makes threats and offers credible commitments in the first place.

D(t) has this tendency given it is under the 45-degree line. This means that any attempt to exercise social power to recapture the lost compliance will be beyond A’s means. It loses those means by virtue of the default. Contrast this with the default G for the following distribution,

While G will mean onlookers begin by assuming all threats/offers costlier than G are momentarily incredible given A will no longer have F(G) to call upon, there is still room for A to cover the cost of more expensive threats/offers rightwards from (G,F(G)) to the 45-degree line if G is only momentary. So A retains the means to build up their resources to reward compliance and punish non-compliance > F(G) and will actively do so if they value maintaining their resource stocks. Those onlookers with thresholds higher than G



will have some reason to believe *that A is of type 1 for their* bargaining game in the moment of observing the default. So once again, threats and offers are often credible, even when there is no danger of a run on A's credibility.[4] The default G may only be momentary, meaning $p < 1$.

The non-cooperative game

So agent A's resources are determined by the sum total of compliance A has secured by way of active threats and offers to reward and punish others. Harsanyi's original formulation of social power appealed to resources as conceptually independent to the game-theoretic analysis. Resources were treated as a primitive. The threshold analysis unpacks the concept to meet the sceptics, but also does justice to its independence by aggregating actions and thresholds independently of their utility and opportunity costs. Yet this is far from saying considerations of game theory are irrelevant to the analysis of social power. In fact, it suggests quite the opposite.

Most individuals will have a certain number of resources F(x) roughly at the y intercept of their threshold graph. I say roughly because in most cases it is actually just after the y intercept, where $x = 1$. This starting point is the cumulative distribution of actions that individuals can secure through resources 'backed' by other agents, such as banks and states. The cost of honouring your threat to the bank and state for failing to deliver on their offers to protect your money and assets is simply to inform others of the bank or state's default, either raising the likelihood of a run on their credibility or a correction and punishment of the individual responsible for the failure to sure off a default. This act of informing is the only action required, hence $x = 1$. We have seen defaults like this trigger runs on the credibility of the state, from the English Civil War, the Orange Revolution in Ukraine, Côte d'Ivoire in 2010, to Venezuela in 2019. So in standard cases we can rely on F(1) to represent the resources of individuals. This would require counting things like votes, legal rights, bank balances, and property portfolios. Wealth can be used to issue credible offers to others by means of commitment mechanisms like legal contracts, backed again by the credibility of the state's threats to punish breaches of contract.

There is something paradoxical about a government who would protect the resource of, say, the vote, but also at the same time be genuinely constrained by it. Yet the paradox can be explained away with the threshold analysis. In the event a politician loses popular support and refuses to hold an election (or ignores the result or rigs the election) there are military officials who are themselves compelled by threats and offers to remove and punish them. There are often institutional safeguards to avoid defaults of this kind. These safeguards are usually enough to secure the compliance of disgruntled politicians. The military officials do their duty not necessarily out of a concern for the state's reputation, but because of the threat of punishment (e.g. a court martial) in the event they do not. Those who have an official duty to punish *them* in the event of non-compliance are subject to similar inducements – and so on and so forth. This network of threats and offers characterizes the state's cumulative distribution of action F(x). State officials may not be directly concerned with the reputation of the state, but there are usually credible inducements *given the state's reputation* to ensure that reputation is preserved with appropriate punishments and rewards. The politician's power grab would only succeed if their non-compliance successfully triggered a credibility cascade.

In democracies no single individual will be in a position to unilaterally choose policymakers with their resources F(1). Groups, however, can do so if they can overcome coordination problems to pool the resources of their members. This is where non-cooperative game theory dominates the analysis. Non-cooperative game theory elaborates a necessary qualification to the resource answer to the question of governance. A group may have limited resources, yet collective action problems might cripple their competition meaning they are free to govern unopposed (Dowding, 2019). These can manifest as simple *n*-person prisoner's dilemmas for their competition, where A > B > C > D for each player. For the simplified 2-person case, this looks like,

	1	2
1	B B	D A
2	A D	**C** **C**

The Nash equilibrium (2,2) is the sub-optimal outcome (given (1,1) > (2,2)) of mutual non-action. Those groups that avoid sub-optimal coordination problems like this can exercise *relatively* more social power than those that do not, despite having relatively fewer resources. If the coordination of resources are in Nash equilibrium, then agents can make stable predictions about the likelihood of coordination given the equilibrium is self-enforcing.

When we couple the identification of these coordination problems with the account of resources developed in the previous section, the problem of commensurability can be squared. If the possibility of sub-optimal Nash equilibria is removed and citizens enjoy stable democratic resources (like free speech, right to office, right to vote, etc.), citizen groups will be able to pool these resources to regulate and replace policymakers. If policymakers can overcome opposition by manipulating the vote (see McKelvey, 1976; Riker, 1988), especially if they have the help of wealthy financial interests, the vote will be a relatively unimportant resource compared with wealth. But coordinated voting can counteract this manipulation (Mackie, 2003). Even if opposition parties are no better, the right to office means it is possible to coordinate to secure better alternatives. In democracies there are credible threats in place that force policymakers to accept electoral results they do not like and tolerate opposition they would rather not. Policymakers do not always comply because of their deep sense of democratic duty, but rather because of the stable and adaptable system of threats and offers F(x) characterizing the state.

Recall Morriss' (2002) objection to the resource approach: a reputation for rewarding or punishing governments at the ballot box will only be so good as there are policymakers who value the group's vote. In terms of the social power bargaining game, however, in well-functioning democracies the threat of electoral coordination by a group should not be construed as directed at the individual policymakers themselves. The above analysis suggests it is better construed as directed at those groups who are supporting those policymakers. These kinds of threats manipulate the disagreement point of the cooperative bargaining game modeling the social power relations between the groups *not* between the group and policymaker. If group A secures policymakers who completely ignore the interests of group B, they face the threat of mobilizing B to replace them. This reciprocal bargaining power would incentivize a compromise between the interests of the two groups. Even if the policymakers themselves are money-driven and have little respect for the broader electorate, they are puppets in the sense they remain in office only so far as they have collective groups supporting them. This is at least the case where the resources of the individuals in those groups are protected from politicians nullifying them.

When the possibility of sub-optimal equilibria is reintroduced, however, it is not at all clear groups will be able to overcome free-riding incentives associated with, say, their right to information (e.g. Downs, 1957), let alone their right to office or free speech, to select and regulate candidates and representatives. This paper suggests the nature and extent of these non-cooperative games is the central empirical question a study investigating governance needs to address. If large voter groups are immobilized in this way, their bargaining disadvantage in the game of social power relative to those groups who are not (arguably the wealthy 0.1%) will be vast The significance of the vote relative to wealth, in other words, is determined by the extent to which large voter groups are crippled by sub-optimal non-action.

Acknowledgements

I would like to thank the editors of JTP, the anonymous referees, Kai Spiekermann, Brad Taylor, Tinghua Yu, Sophie Fehlberg, and participants at the POTY workshop at the London School of Economics in 2019 for help. I am especially grateful to Keith Dowding who I suspect disagrees with virtually every sentence in the paper, but whose work and encouragement inspired me to think about the issues.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

William Bosworth https://orcid.org/0000-0001-5234-7717

Notes

1. Weber's definition follows "common usage" (Barry, 2002: 161) and captures what Morriss (2002: 40) calls the second "evaluative context" of power-talk, which is the same context Dahl (2005) [1961]: 1-3) framed the question of governance in. To insist on it as the only correct definition of 'social power', however, would just fuel verbal disputes (Bosworth, 2020: 306-8).
2. The infinitely repeated bargaining game with a fixed bargaining cost or discount factor can also capture similar dynamics (Rubinstein, 1982). The costs associated with each round could be interpreted as an incremental punishment whose rate is fixed by a threat made at the beginning of the game. The setup would nevertheless still rely on each punishment being credible, which cannot be assumed.
3. That players have consistent common priors is perhaps unintuitive, but part of the revelation of Bayesian game theory is we have a number of proofs to that effect, namely that "any Bayesian game with finite type sets is equivalent to a Bayesian game with consistent beliefs" (Myerson, 1991: 73).
4. This suggests flaking will be more likely for rewards and punishment that require high levels of compliance (such as wars) but also that it is rational to take such threats and offers seriously. Empirical verification here would be a good test for the theory overall.

References

Arendt H (1954) *Between Past and Future*. New York: Viking Press.

Barry B (1989a) *Democracy, Power and Justice: Essays in Political Theory*. Oxford: Clarendon Press.

Barry B (2002) Capitalists rule OK? Some puzzles about power. *Politics, Philosophy and Economics* 1(2): 155–184.

Barry B (2003) Capitalists rule OK? A commentary on Keith Dowding. *Politics, Philosophy and Economics* 2(3): 323–341.

Bosworth W (2020) An interpretation of political argument. *European Journal of Political Theory* 19(3): 293–313.

Connolly WE (1974) *Terms of Political Discourse*. Princeton, New Jersey: Princeton University Press.

Dahl R (1957) The concept of power. *Behavioral Science* 2(3): 201–215.

Dahl R. (2005) [1961] *Who Governs? Democracy and Power in an American City*. New Haven: Yale University Press.

Dowding K (1996) *Power*. Buckingham: Open University Press.

Dowding K (2019) *Rational Choice and Political Power*, 2nd Edition Bristol: Bristol University Press.

Downs A (1957) *An Economic Theory of Democracy*. New York: Harper & Row.

Fudenberg D and Maskin E (1986) The folk theorem in repeated games with discounting or with incomplete information. *Econometrica* 54(3): 533–554.

Harsanyi JC (1962a) Measurement of social power, opportunity costs, and the theory of Two-person bargaining games. *Behavioral Science* 7(1): 67–80.

Harsanyi J (1962b) Measurement of social power in n-person reciprocal power situations. *Behavioral Science* 7(1): 81–91.

Harsanyi J (1967) Games with incomplete information played by “Bayesian” players. Part I. The basic model. *Management Science* 14(3): 159–182.

Kreps DM and Wilson R (1982) Reputation and imperfect information. *Journal of Economic Theory* 27(2): 253–279.

Lukes S (2005) *Power: A Radical View*, 2nd Edition London: Palgrave Macmillan.

Mackie G (2003) *Democracy Defended*. Cambridge: Cambridge University Press.

McKelvey RD (1976) Intransitivities in multidimensional voting models and some implications for agenda control. *Journal of Economic Theory* 12(3): 472–482.

Morriss P (2002) *Power: A Philosophical Analysis*, 2nd Edition Manchester: Manchester University Press.

Myerson RB (1991) *Game Theory: Analysis of Conflict, Harvard University Press*. Cambridge: Mass.

Nash J (1953) Two-person cooperative games. *Journal of the Econometric Society* 21(1): 128–140.

Olson M (1965) *The Logic of Collective Action: Public Goods and the Theory of Groups*. Cambridge: Harvard University Press.

Petley C (2018) *White Fury: A Jamaican Slaveholder and the Age of Revolution*. Oxford: Oxford University Press.

Pettit P (1997) *Republicanism*. Oxford: Clarendon Press.

Polsby N (1980) *Community Power and Political Theory*. New Haven: Yale University Press.

Riker WH (1988) *Liberalism Against Populism: A Confrontation Between the Theory of Democracy and the Theory of Social Choice*. Long Grove, Il: Waveland.

Rubinstein A (1982) Perfect equilibrium in a bargaining model. *Econometrica* 50(1): 97–109.

Schelling T (1978) *The Strategy of Conflict*. Cambridge, Mass: Harvard University Press.

Selten R (1978) *The Chainstore Paradox. Theory and Decision*, 9: 127-159.

Weber M (1978) *Economy and Society*. California: University of California Press.

Wollstonecraft M (1993) *A Vindication of the Rights of Woman, A Vindication of the Rights of Men*. Oxford: Oxford World Classics.

Wolfinnger R. E. (1971) Nondecisions and the Study of Local Politics. *American Political Science Review*, 65: 1063–1080.