

TESTING DESIGN HYPOTHESES: USING FORMAL MODELS TO TEST INSTITUTIONAL DESIGN PROPOSALS

David Wiens

Philosophers have not been shy to prescribe institutional reforms under the rubric of nonideal theory. This state of play seems natural to most normative philosophers. Since political and social institutions are among the primary concerns of normative political philosophers, they will naturally have something about the way these institutions *should be* designed. To the extent that they care about improving current social conditions, those prescriptions will be intended as feasible solutions to various current injustices.

Treating institutional design as a wholly normative philosophical enterprise looks suspicious from the view of social scientists. The problem, in their view, is that philosophers have often shirked a responsibility to present evidence for the feasibility and effectiveness of their proposals. We often seem unwilling or unable to offer little else beyond our intuition. Unfortunately, intuition is an unreliable guide here, informed as it often is by unjustified optimism (or pessimism).

To begin to address this problem, I argue that each institutional design prescription should be treated as the counterfactual hypothesis that the proposed institution is feasible and would, if implemented, effectively bring about moral progress. *Qua* hypotheses, design proposals are answerable to a *testing requirement*: design proposals bear a burden of showing that their associated hypotheses are plausibly true. A shortage of useful tools for analyzing counterfactual scenarios makes this testing requirement difficult to meet. I argue that formal game theoretic models can partly fill this void. Formal modelling has two key advantages that lend it to serving the required testing role. First, models can be useful for isolating causal mechanisms and facilitating their close investigation. This is important since institutional design primarily aims to alter existing social causal processes with the hope of improving the outcome we observe. Second, models can be useful for examining the strategic dynamic of counterfactual worlds, which is important given that design proposals are hypotheses about counterfactual worlds. Together, these advantages permit us to examine the joint logical implications of a set of premises about

AUTHOR'S NOTE. The model in this paper was presented to an audience at the University of Michigan. Thanks to participants for useful discussion. Thanks also to Bill Clark, Dan Little, Skip Lupia, and Jim Morrow for extensive feedback on early drafts of various portions of the paper.

the strategic interactions that would arise if the proposed institutions were implemented. This constitutes a useful test.

1. HOW TO THINK ABOUT DESIGN PROPOSALS

To frame my discussion, consider the following examples of institutional design prescriptions.

- (1) The international community should adopt a standard of recognitional legitimacy whereby a state must meet four explicitly moral criteria to be recognized as legitimately sovereign by other sovereign states: (i) the Internal Justice Condition; (ii) the External Justice Condition; (iii) the Nonusurpation Condition; and (iv) the Minimal Democracy Condition.¹
- (2) Fledgling democracies should enact an *Odious Debt Amendment*, i.e., a constitutional amendment to require that debts incurred by future unconstitutional governments (i.e., governments who acquire power by unconstitutional means) not be serviced at public expense.²
- (3) States should adopt a ‘trust-and-tariff’ policy to regulate their trade with states who purchase natural resources from the worst autocratic governments (i.e., defector states).³

How should we view these proposals? Here are three options. For any institutional design proposal *P*,

- *P* could express an ideal to which we should aspire;
- *P* could constitute part of an analysis of a concept;
- *P* could prescribe a feasible solution to actual injustice(s).

These options aren’t mutually exclusive. But each kind has different criteria that *P* must meet for *P* to be a successful instance of that kind. Alas, the scope of this paper is limited to an elaboration of the third kind.

What criteria must *P* meet to successfully prescribe a feasible solution to actual injustice? At a minimum, it must be true that (1) *P* is *feasible* and, (2) *P* is a candidate

1 On the content of these conditions, see Buchanan 2004, pp. 269–272, 275f, 278f.

2 Pogge 2002, p. #.

3 For the details of the ‘trust-and-tariff’ proposal, see Wenar 2008, esp. §10.

solution to the target problem. How do we determine whether P meets these criteria? Drawing on some work I have done elsewhere, I suggest that we treat design prescriptions as a sort of hypothesis.⁴ In particular, we should treat some proposal P as the design hypothesis that *the prescribed institution (i) can be established given the constraints on institutional establishment and (ii) will successfully achieve its objective under the conditions in which it will be required to operate*. The fact that P *qua* hypothesis must be true to count as a successful recommendation of a feasible solution generates a *testing requirement*. That is, proposals of this sort bear the burden of showing that the associated hypothesis is true. But how do we test an hypothesis of this sort? How do we know whether an hypothesis such as this is true?

To simplify the analysis, I decline to discuss the feasibility criterion here and restrict my attention to an analysis of the solution criterion.⁵ As a first pass, we can standardize our analysis of P as follows.

(P) Institution I is more likely than not to bring about outcome O (under the relevant conditions).⁶

This isn't quite satisfactory, though. Unless we know what 'is more likely than not to' means, it's still not clear how to go about testing P . Fortunately, the resources for a straightforward analysis of the notion of likelihood already exist: we can analyze likelihood in terms of possible worlds. The basic idea is this. Define a 'possible world' as '[a] complete [way] of how reality might have been'.⁷ Define the set of 'nearby I -worlds' as those possible worlds that are identical to the actual world in all ways except the following:

⁴ Wiens 2010.

⁵ Here's a sketch of my provisional analysis of feasibility. Define the 'closest F -world(s)' as that (those) possible world(s) that hold fixed the following features of the actual world: (1) history up to the present time t ; (2) the salient constraints on institutional establishment (e.g., the relevant agents' cognitive and motivational biases, incentive structure, computational and technological limitations); and (3) the social and political processes by which institutions are established. (This last clause is supposed to exclude worlds wherein institutions are established spontaneously or by, say, divine command.) Using this notion of 'closest F -world', we can define 'feasibility' as follows.

An institutional design proposal P is *feasible* just in case there exists at least one closest F -world where the proposed institution I is established at some time after t .

In lieu of further discussion of feasibility, I simply point the reader to some of the literature on feasibility with which I'm familiar: Brennan and Pettit (2005); Cowen (2007); Jensen (2009); Raikka (1998); Reddy (2005).

⁶ The parenthetical clause restricts our attention to the salient context, viz., the set of conditions under which I must be established and subsequently operate.

⁷ Melia 2003, 18.

- (1) *The fact of establishment.* The proposed institution *I* is established in the nearby *I*-worlds but not in the actual world;
- (2) *The mutatis mutandis clause.* The nearby *I*-worlds exhibit whatever departures must be made from the actual world to make it true that *I* is established.

The latter clause leaves open the possibility that there might be multiple nearby *I*-worlds, since there might be several distinct causal paths by which *I* is established, each of which entails distinct departures from the actual world. Thus, nearby *I*-worlds are individuated by their departures from the actual world.

Using the possible worlds apparatus, we can now reformulate the analysis of *P* as follows.

(***P****) *O* is realized in a sufficient number of nearby *I*-worlds.

It will be important to define the sufficiency threshold. We can say for sure that *O* must be realized in more than half of the nearby *I*-worlds, since *I* is supposed to be more likely *than not* to bring about *O*. How many more than half is still an open question, but one that we can safely leave vague for now.

To show how my analysis works in practice, let's reformulate the aforementioned proposals in accordance with (***P****), now including the target outcome of each proposal.

- (1*) The human rights performance of sovereign states is improved in a sufficient number of nearby worlds wherein the international community adopts of a standard of recognitional legitimacy whereby a state must meet four explicitly moral criteria to be recognized as legitimately sovereign by other sovereign states.
- (2*) The number of coups is reduced and democratic reforms are consolidated in a sufficient number of nearby worlds wherein fledgling democracies enact a constitutional amendment to require that debts incurred by future unconstitutional governments not be serviced at public expense.
- (3*) Citizens' resource rights are respected in a sufficient number of nearby worlds wherein states adopt a 'trust-and-tariff' policy to regulate their trade with states who purchase natural resources from the worst autocratic governments.

It will be important for my discussion in the last section to notice that each of these is an hypothesis about the outcome that would be caused by a particular institutional arrangement were that arrangement to be established. In other words, each of these is

an hypothesis about *the outcome that results from the operation of a particular causal mechanism in a set of counterfactual worlds.*

2. HOW TO THINK ABOUT FORMAL MODELS

How can institutional design prescriptions *qua* counterfactual hypotheses meet the burden of showing that they are plausibly true? The problem here is that hypotheses about counterfactuals can't be tested in the same way as hypotheses about the actual world. The central claim of this paper is that formal game theoretic models can be useful for testing design hypotheses. But before I undertake to argue for that claim, I make a pitstop to enumerate several key considerations concerning formal models on which my argument will rely.

2.1. *Models as Isolating Devices*

To generate the intuition behind the idea of a formal model as an isolating device, consider the use of material controls in laboratory experiments. The objective of experimental controls is to eliminate causal noise and isolate the causal relationship of interest for more detailed investigation. Similarly, formal models are 'thought experiments' that control for causal noise through the use of idealizing or exaggerating assumptions with the aim of isolating the causal relationship of interest and placing it in an environment where the causal connection is stable.⁸ As such, models are fundamentally *representations* of the target system, in at least two senses. The first sense is that a model *represents*, or 'stands in place of', the target system as a subject of inquiry.⁹ Attempts to analyze the operation of social causal mechanisms¹⁰ by investigating the target system directly are unlikely to succeed given the complexity of the social causal structure of our world. Models can aid us in our inquiry by analytically pulling apart causal interactions and enabling us to move some parts of the model while holding others fixed, an exercise we are typically unable to do with the mechanisms as we encounter them in the world. Hence, the turn to investigating model worlds is one of necessity. Without epistemically reliable 'direct' access to the causally-complex actual world, we're limited to investigating model worlds. But if a model world adequately represents the actual world, then we have some basis for drawing inferences about the actual world from our examination of the model.

8 Cf. Mäki 2005.

9 Cf. Mäki 2009.

10 For detailed discussion of social causal mechanisms, see the Hedström and Swedberg, Schelling, Elster, and Cowan contributions to Hedström and Swedberg 1998, as well as Elster 2007, pt. 1.

The second sense in which a model represents the target system addresses this adequacy requirement: a model must *resemble* the target in certain salient ways. In particular, if we intend to draw inferences about the actual world from our investigation of the model, the core of the model must consist of well-confirmed causal regularities in the target system. In addition, the mechanism doing the work in the model must be a mechanism that is in operation in the target system.¹¹ If the model fails to resemble the target system in important ways, then we are unable to make reliable inferences from the model to the target.

2.2. *Models as Credible Counterfactual Worlds*

Contra the isolation view, Robert Sugden argues that models are counterfactual worlds constructed from the ground up. Using Akerlof's (1970) 'market for lemons' model and Schelling's (1978) 'checkerboard' model of racial segregation as examples, Sugden highlights several features of the practice of formal modelling that seem inconsistent with the isolation view. First, many formal models are not built upon well-confirmed empirical features of the real world, but are built on substantive (usually false) assumptions, which do most of the work in generating the results of the model. Many formal modellers do not aim to isolate causal mechanisms known to operate in our world, but seek to investigate the causal structure of a hypothetical model world. Second, although modellers think we can learn something about the actual world from their models, many do not think the link is direct or deductive. In particular, many modellers decline to suggest testable causal hypotheses about the actual world. Instead, the argument seems to be by analogy: the argument from model to target system identifies relevant similarities between the two worlds and then suggests that the best explanation for the similarity in outcomes is a similarity in causal structure.

Consequently, on Sugden's view, models do not serve as 'surrogate systems' in the sense sketched above. In investigating a model world, we are simply deducing claims about a counterfactual world, a world that *could have been* actual. To draw conclusions about the actual world from the model, we need to make inductive inferences from model to actual world. The question, then, is how such inductive inferences are justified. Sugden claims that the more the model world resembles the actual world in the relevant respects, the more confident we are in inferring things about the causal structure of the actual

¹¹ If we're using a model to propose a theory to explain some empirical regularity, then we might not know in advance whether the mechanism at work in the model is at work in the target system. In this case, the mechanism at work in the model must be *a candidate* for operation in the target system. We then aim to substantiate this conjecture by empirically testing the observable implications of the proposed theory.

world from the causal structure of the model world. This is where the notion of *credibility* is supposed to do its work. The more credible a model is as a *candidate* for truth, the more justified we are in drawing model-to-target system inferences.

A model's credibility turns on three considerations. First, a model's credibility increases as the distance between the counterfactual world it describes and the actual world decreases. We must in some way think that the world described by the model is 'close' to ours in important respects. Thus, as with the isolation view, the issue of resemblance arises when making inferences from model to target system. Second, the description given by the model should be internally coherent. The actions that generate the outcome must be consistent with what we know about the behaviour of agents in the model world. Third, important features of the model world must cohere with what we know about the actual world. That is, key features of the model world must be representative of the actual world, even if they are not replicable. Take, for instance, the *homo economicus* character that shows up in so many models. It's patently false that our world is entirely populated by such characters, and the capacities attributed to this species in model worlds are often unlike (sometimes radically so) the capacities of the species in this world they are supposed to represent. Nonetheless, we can recognize key features of *homo economicus* in *homo sapiens*; the former is coherent as a stylized representation of the latter, even if the former is not, strictly speaking, replicable.

2.3. *Models as Arguments*

Whether a model is best viewed as an isolating device or as a credible counterfactual world, the question of how we can learn something about the actual world from a model that makes deliberately false assumptions lingers.¹² To begin to answer this question, note that, whatever the differences between the two views, inferences drawn from a model (on either view) are inferences from premises to a (set of) conclusion(s).¹³ In formalizing interactions between strategic agents, formal game-theoretic models enable us to clarify the causal interconnections between microlevel behaviour and macrolevel outcomes and investigate the validity of the inferences from premises about the characteristics of agents

¹² There has been much debate concerning the similarities and differences between these two views. I don't aim to contribute to this debate, but see, e.g., *Erkenntnis* vol. 70, no. 1 (2009), especially the papers by Knuttilla, Kuorikoski and Lehtinen, and Mäki. Personally, I don't think we need to decide. Some models are clearly attempts to isolate (e.g., Bueno de Mesquita et al.'s (2003) 'selectorate' model; Fearon and Laitin's (1996) model of interethnic cooperation); Sugden provides examples of models that are clearly counterfactual worlds built from scratch. It seems to me that the character of the model is determined by the argumentative purpose of the model, in a different sense than the one that I elaborate in the remainder of this section.

¹³ Cf. Kuorikoski and Lehtinen 2009, 122.

and their interactions to conclusions about social phenomena. The advantage of formal modelling as a method for investigating explanations of social outcomes is that it enables us to formalize and examine the dynamics of an interconnected system of several moving parts and investigate causal claims by moving the variable of interest while holding the others fixed.

As Kuorikoski and Lehtinen point out, viewing formal models this implies that they do not pose any special epistemological problem of inference. This closes the ‘inductive gap’ between the model world and the actual world implied by the use of idealizing — indeed, false — assumptions. As with any argument, if the initial premises are true and the inferences valid, then the conclusions are sound. This implies that ‘all epistemic questions about modelling can be conceived as concerning either the reliability of the assumptions or the reliability of the inferences made from them.’¹⁴

Of course, all models include some false premises! So long as this practice continues, what can a model that employs unrealistic assumptions tell us about the actual world? Whether and which false assumptions are problematic depends on the argumentative purpose of the model. If the modeller aims to argue for some conditional claim (‘If x and y , then O ’), to demonstrate the conditions under which some outcome is possible, or to undermine an impossibility claim (‘ A says O is impossible, but this model shows that O can occur when x and y ’), then he needs nothing more than to investigate the causal structure of a (counterfactual) model world. In these cases, there is no problem posed by using false assumptions.

If the modeller aims to tell us something about the actual world, the model need not resemble the actual world in all respects. It need only resemble the actual world in the *salient* respects, where ‘salient’ is determined by the argumentative context in which the model is situated. Models typically aim to say something about the operation of a specific causal mechanism. Hence, the mechanism in the model must closely resemble the mechanism operating in the actual world. (Or, if the actually operative mechanism is unknown, the model mechanism must at least resemble a candidate for the mechanism operating in the actual world.) In other words, the *substantive premises about the mechanism in question* must be plausibly true of (or plausible candidates for truth regarding) the actual mechanism. But our *auxiliary* assumptions — assumptions such as convex strategy spaces, standardizing payoffs to range from 0 to 1, geometric time discounting — often won’t or can’t be true; we frequently need to make these unrealistic assumptions to make the model mathematically tractable. What we need, then, is

¹⁴ Kuorikoski and Lehtinen 2009, 120.

some way to determine whether the results of the model are driven by our substantive assumptions or our auxiliary assumptions.

Kuorikoski and Lehtinen argue that the way to do this is via *robustness analysis*.¹⁵ The basic idea is to investigate the model under different sets of auxiliary hypotheses while holding the substantive assumptions fixed. Since auxiliary hypotheses are usually false for good (modelling) reasons, the prescribed changes among sets of auxiliary hypotheses will not be from false assumptions to true ones, but changes from one set of false assumptions to another. For example, to investigate whether the results of the selectorate model are driven by the dynamics of leader selection institutions or by (e.g.,) the form of the utility functions, we investigate several versions of the model holding the causal mechanism fixed while changing the form of the utility functions. If we find that the model generates fundamentally similar results across different sets of auxiliary hypotheses, we'll be more confident that the results are indeed driven by the mechanism under investigation and not errors in the auxiliary hypotheses. If our substantive assumptions are approximately true of the world, then the robustness of the results to changes in auxiliary hypotheses gives us reason to think that the model illuminates the operation of the mechanism in the actual world.

There are three main points to take with us from this section to the next. First, models can be useful for isolating causal mechanisms and facilitating their close investigation. Second, models can be useful for examining the strategic dynamic of counterfactual worlds. Third, whether and which false assumptions pose a problem depends on the argument the model aims to make. Only a model's substantive assumptions — i.e., the premises about the core causal mechanism — must be true. The model's argumentative purpose determines which assumptions are substantive and which are auxiliary.

3. HOW FORMAL MODELS CAN HELP US THINK ABOUT DESIGN PROPOSALS

We should view any institutional design prescription P as the hypothesis that the prescribed institution I is feasible and is more likely than not to bring about some morally desirable outcome O . Such hypotheses must meet a testing requirement. This requires that we answer at least the following two questions.

- (1) Is I *feasible*?
- (2) If established, is I *capable of bringing about* O ?

¹⁵ Kuorikoski and Lehtinen 2009.

(1) asks about the causal structure of the actual world; in particular, it asks about the operation of particular causal mechanisms that might facilitate or impede the establishment of *I* in the actual world. Since I haven't offered a detailed analysis of feasibility here, I leave a discussion of how models could help answer (1) for another time.¹⁶

(2) asks about the causal structure of counterfactual worlds, viz., of the nearby worlds wherein the proposed institution is established. In particular, to answer this question, we must investigate the effects of *I qua* causal mechanism in those counterfactual worlds. Again, since models are useful for isolating particular causal mechanisms for close examination and they are useful for investigating counterfactual worlds, we can use models to answer (2). Since we are asking about the effect of a mechanism that is not operative in the actual world, it looks like the ontology of the actual world won't constrain the model's substantive premises in the way that they are constrained in answering the feasibility question. But if we're to have any confidence that *I* is more likely than not to bring about *O in the actual world*, the model's substantive premises should be somewhat constrained by the ontology of the actual world. Unfortunately, it's not clear what to say here. We could say that the model's substantive premises be roughly true of how *I could* or 'would' operate in the actual world. But what we'll mean by these modal terms is 'how *I* operates in (a) nearby *I*-world(s)', which is to say that the ontological constraints are provided by the counterfactual world rather than the actual world. Instead, we should turn to the *mutatis mutandis* clause to provide the salient constraints. A nearby *I*-world departs from the actual world only in the respects that are necessary to establish *I*. Thus, we could say that the model's substantive premises must be true of the actual world, *mutatis mutandis*. At the very least, the operation of *I* in the model should be recognizable as a mechanism that could plausibly operate in the actual world; in Sugden's terminology, they must be 'credible candidates for truth' in the actual world.

4. AN EXAMPLE

The foregoing is likely too abstract to offer a tangible idea of how we might use models to test design hypotheses. Thus, to illustrate, I present a model that tests whether Pogge's

¹⁶ But here's a sketch. We should answer this question by investigating whether *I* is established in any of the closest *F*-worlds. Recall that the closest *F*-worlds are counterfactual worlds. Models are useful for isolating particular causal mechanisms for close examination and they are useful for investigating counterfactual worlds. So long as the model's premises concerning the causal mechanism(s) under investigation are (roughly) true of those mechanisms in the actual world, our conclusions about the operation of these mechanisms in the model—in particular, whether they facilitate or inhibit the establishment of *I*—can ground conclusions about the feasibility of *I* in the actual world.

proposed *Odious Debt Amendment* is capable of bringing about the anticipated outcome. Recall Pogge's design hypothesis (as I've formulated it):

The number of coups is reduced and democratic reforms are consolidated in a sufficient number of nearby worlds wherein fledgling democracies enact a constitutional amendment to require that debts incurred by future unconstitutional governments not be serviced at public expense.

To test this, I construct a model of a counterfactual world wherein the proposed amendment is adopted. I then investigate the effect of the amendment on the incidence of coups. I assume that the amendment is credibly enforced in this world. This is an important assumption, as it presents a best-case scenario for Pogge's proposal. Even if Pogge's proposal could be established, it might still be plagued by enforcement issues, rendering it ineffective. The investigation here assumes away this problem, examining whether the amendment would effectively deter coups even if enforcement were unproblematic.

Since Pogge's claim is a comparative one — viz., that the amendment would *reduce* the number of coups — I must also construct a model of a world where the amendment is not in force (i.e., the actual world) and use the incidence of coups in that world as a baseline for comparison. If the incidence of coups in the amendment world is lower than the incidence of coups in the non-amendment world, then we have some evidence that Pogge's design hypothesis — at least with respect to the solution criterion — is true.

4.1. *The Model*

For the purpose of testing Pogge's proposal, we're particularly interested in the interaction between two players, an *autocrat* and a *lender*.¹⁷ (To distinguish between the two throughout, I use male pronouns for the autocrat and female pronouns for the lender.) We're interested in what happens once a fledgling democratic government is in place. So assume that our investigation starts with a fledgling democratic government is in place. At the beginning of the game, the autocrat is only a potential autocrat; he poses a threat to undertake a coup against the democratic government. If the autocrat attempts to seize power, I assume that he succeeds and establishes an autocratic government. This means that the autocrat is deterred only by what happens after he acquires power. This assumption helps separate the effect of receiving a loan (or not) on the autocrat's decision from the effect of uncertainty about his chances of successfully acquiring power. This is useful since it is the effect of the loan in which we're interested here.

¹⁷ I leave most of the formal analysis of the model to an appendix.

The lender is a foreign creditor in the business of extending loans to foreign governments. I assume that the lender is the autocrat's only available source of credit. This assumption simplifies the analysis of the model by reducing the number of actors we need to keep track of, but it gives the lender a monopoly on borrowing opportunities, making it an unrealistic assumption. I address this worry below.

The interaction proceeds according to the following timeline.

- (1) The players observe whether the amendment is in force or not.
- (2) The autocrat chooses whether to seize power. If he seizes power, the game moves to phase 3. If he refrains, the game ends with the status quo in place. The autocrat gets a payoff of 0 and the lender gets a payoff of 1.
- (3) The autocrat chooses whether to request a loan.
- (4) The lender chooses whether to grant a loan to the autocrat.
- (5) The autocrat spends the optimal amount of total income in an attempt to maintain political support.
- (6) There is an exogenous challenge to the autocrat's power. The autocrat remains in office with probability $p(c)$ and is ousted with probability $1 - p(c)$. If the autocrat remains in power and received a loan, he repays the loan with probability $1 - \gamma$.¹⁸ If he is ousted, he pays a fixed cost k . If the lender lent to the autocrat and the amendment is not in force, the autocrat's successor repays the loan. The players receive their payoffs.

The payoffs are as follows. Following Pogge, I assume that coups are motivated by 'greed' rather than 'grievance'.¹⁹ Accordingly, the autocrat's payoff for any outcome is an increasing function of the total amount of revenue at his disposal and the probability he remains in office and a decreasing function of the cost of maintaining political support — i.e., the amount of the revenue he must spend to maintain political support — the interest rate on the loan, and the fixed cost of being ousted. The autocrat's revenue is the sum of two revenue sources. One is the extra-credit revenue function, which is a function from non-loan income sources, such as resource extraction, to revenue. The other is the credit revenue function, which is a function from the amount of the loan to revenue. I assume that credit revenue is instantaneously realized; that is, current-period

¹⁸ γ is defined below.

¹⁹ Cf. Collier and Hoeffler 1998; Collier and Hoeffler 2004.

credit revenue is generated from current-period loan income. (For symmetry, I assume that extra-credit revenue is also instantaneously realized.) This assumption serves the same purpose as the assumption that the autocrat is successful if he undertakes a coup. It insures that the autocrat realizes the gains of acquiring power upon undertaking a coup, thereby making his decision rest solely on the benefits of seizing power. This is important because the focus of our investigation is the effect of reducing the benefits of seizing power on the autocrat's decision to undertake a coup. I also assume that the credit revenue function is increasing and concave; that is, loan income yields diminishing marginal revenue. For simplicity, I treat the amount of the loan as fixed. The autocrat either receives a loan or doesn't. If the autocrat receives a loan, he must repay the lender the amount of the loan plus interest $(1 + r)$, with $r > 0$. The revenue generated by the credit revenue function can be less than, equal to, or greater than $1 + r$. If the revenue generated from the loan is no greater than $1 + r$, I call the autocrat 'unproductive'. If the loan revenue is strictly greater than $1 + r$, I call the autocrat 'productive'. An unproductive autocrat parlays the loan income into a net revenue decrease or (at best) net revenue stagnation. A productive autocrat parlays the loan income into net revenue growth.

Once the autocrat seizes power, his objective is to choose a spending level c that maximizes his payoff given the amount of loan revenue he receives. Since the autocrat receives a payoff of 0 if he refrains from seizing power, he attempts a coup if and only if his payoff to doing so, given his expected net income, is greater than 0.

Once he's spent the chosen amount, the autocrat faces a challenge and stays in power with probability $p(c)$, which is a function of the amount $c > 0$ that he spends on maintaining political support. The autocrat can't spend more than the net revenue at his disposal after paying back the loan; thus, the maximum c is the difference between the autocrat's total revenue and the amount he must repay on the loan. I assume that $p(c)$ is an increasing and concave function; i.e., c yields diminishing marginal probability. I also assume that if the autocrat spends nothing, he is guaranteed to be ousted [$p(0) = 0$]. Further, even if he spends everything, I assume the autocrat can't guarantee his political survival [$p(M) < 1$]. Finally, spending everything yields the maximum political survival probability.

I assume that the autocrat spends the optimal amount once in office. This means that he maximizes his expected payoff in office, thereby making the benefits to seizing power as large as possible. There are two distinct optimal spending levels. If the autocrat obtains a loan, I denote the optimal spending level c^* . If the autocrat does not obtain a loan, I denote the optimal spending level c^0 . Since the autocrat can't credibly commit to spending a suboptimal amount once in office, the optimal c is fixed by the definitions

of c^* or c^0 .²⁰ Once the optimal spending level is fixed, this fixes $p(c)$. Thus, since the autocrat (and the lender) knows in advance how much total revenue would be at his disposal were he to receive a loan once in office, he (and the lender) can calculate in advance how much he will spend on maintaining political survival and, thereby, the probability that he will survive a challenge.

I assume that the lender's decision to lend or not is motivated solely by expected profit. If the lender chooses not to lend, she keeps her money and receives a payoff of 1. Whether the amendment is in force matters to the lender. If she lends and the amendment is not enacted, then the lender is able to extract repayment from the autocrat's successor. If she lends and the amendment is enacted, then she cannot extract repayment from the autocrat's successor and loses her money if the autocrat is removed from office. I assume that all debtors default with probability γ . If the amendment is not in force, the lender's payoff for lending is an increasing function of the interest rate and a decreasing function of γ . If the amendment is in force, the autocrat's survival probability matters; now the lender's payoff for lending is an increasing function of the interest rate and the autocrat's survival probability and a decreasing function of γ .

Although I assume that the lender is the autocrat's only available source of credit, I can capture the effect of competition among multiple lenders by assuming that the lender is a 'price-taker' and that competition drives the market interest rate down to the minimum acceptable rate for all lenders given γ . Any lender lends if and only if doing so yields a greater payoff than not lending. Since this is true when $(1 - \gamma)(1 + r) + \gamma 0 \geq 1$, the market rate is set to $\underline{r} = \frac{\gamma}{1 - \gamma}$.

4.2. *No Amendment*

I now characterize the outcome when the amendment is not in place. This serves as the baseline for assessing Pogge's claims about the effect of switching to a world where the amendment is enacted. In the base model, I assume that the autocrat has no special difficulty borrowing money, so he borrows on the same terms as any other candidate for a loan. Consequently, given the market interest rate, the lender always lends to the autocrat.

For the purpose of assessing Pogge's prediction about the amendment's effect, it doesn't matter how well-off the autocrat is when in power — that is, whether he gets a loan (or not) when he prefers receiving one (or not). Consequently, I don't try to determine whether the autocrat requests a loan or not in equilibrium. More generally, I don't solve

²⁰ These definitions are given in the appendix.

for the game's equilibrium — for the set of actions that constitute each player's best replies to the other player's actions under all contingencies. Instead, I solve for the autocrat's *seize threshold* for both cases, loan and no loan. These thresholds identify the survival probability where the autocrat is indifferent between seizing power and refraining for both cases. For all probabilities greater than or equal to a threshold π , the autocrat seizes power. Define the autocrat's *coup space* as the range of probabilities from a threshold π to 1 inclusive. Pogge's claim is that the amendment will reduce the autocrat's coup space. Thus, to test Pogge's claim, the relevant investigation compares the size of the coup space in the no amendment world to that in the amendment world. In this section, I characterize the coup space in the no amendment world. In the next section, I characterize the effect of the enacting the amendment on the size the coup space.

The autocrat can be in one of two circumstances: he can either receive a loan or not. If the autocrat receives a loan, then the seize threshold is π^* . If the autocrat does not receive a loan, then the seize threshold is π^0 .²¹ The relationship of these cutpoints to each other differs depending on whether the autocrat is productive or unproductive.

Figure 1 depicts the key results of this section by showing the coup space for both types of autocrat. The top line shows that the loan threshold is no lower than the no loan threshold if the autocrat is unproductive. The bottom line shows that the no loan threshold is higher than the loan threshold if the autocrat is productive. Note that the location of π^0 and π^* along the interval is not important. π^0 is in the same location for each type because it's not a function of the autocrat's productivity, whereas π^* is a decreasing function of the autocrat's productivity. The point here is to illustrate the location of the two thresholds *relative to each other*. The key question now is how introducing the amendment affects these coup spaces.

4.3. Amendment

As is the case without the amendment, the lender lends if and only if doing so yields a greater payoff than not lending. Define \hat{r} as the minimum acceptable rate for lending to the autocrat given the amendment. Recall that the market rate is \underline{r} . Note further that $\underline{r} < \hat{r}$ regardless of the autocrat's survival probability. Since the autocrat could agree to borrow at a rate greater than \underline{r} if doing so benefited him, the interest rate could be greater than, equal to, or less than \hat{r} in the amendment world. However, the autocrat wouldn't agree to borrow at a rate greater than \hat{r} , since doing so harms him unnecessarily. Thus, in the

²¹ The formal definitions of these thresholds are given in the appendix.

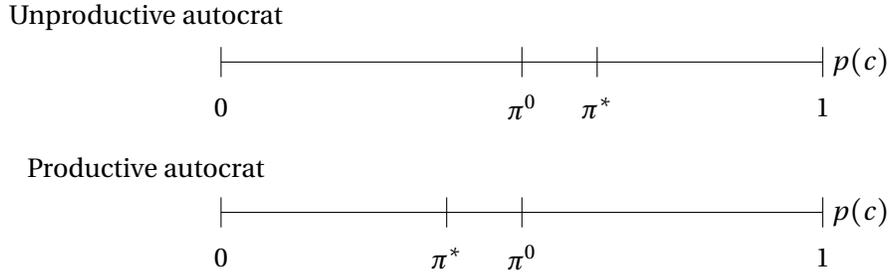


Figure 1. Relative thresholds for unproductive and productive types without the amendment

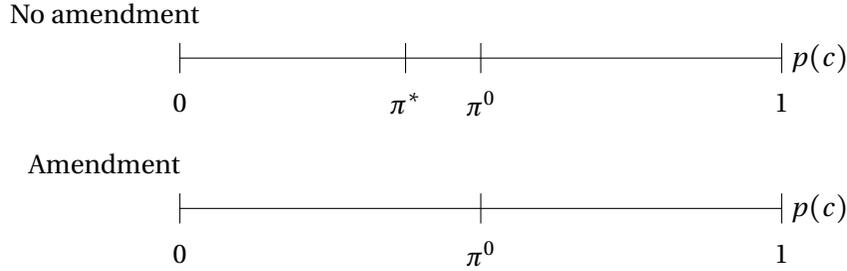


Figure 2. Threshold change for productive type without loan

amendment world, the interest rate ranges from \underline{r} to \hat{r} . This means that the lender could either lend to the autocrat or not, depending on the interest rate.

Suppose for now that the lender doesn't lend. In this case, the effect of the amendment is fairly straightforward. When there was no amendment in force, the lower bound of the autocrat's coup space was defined by the lower of π^0 and π^* . When the lender doesn't lend, the effect of the amendment is to remove the option of seizing power and receiving a loan. Accordingly, with the amendment in place, the lower bound of the autocrat's coup space is now defined by the location of π^0 . From fig. 1, we see that that the size of an unproductive autocrat's coup space is unchanged, since $\pi^0 \leq \pi^*$. However, we also see that the amendment reduces a productive autocrat's coup space. Now that loans are no longer available, π^* is no longer relevant. This is depicted in fig. 2.

Now suppose that the lender lends to the autocrat; that is, suppose the interest rate

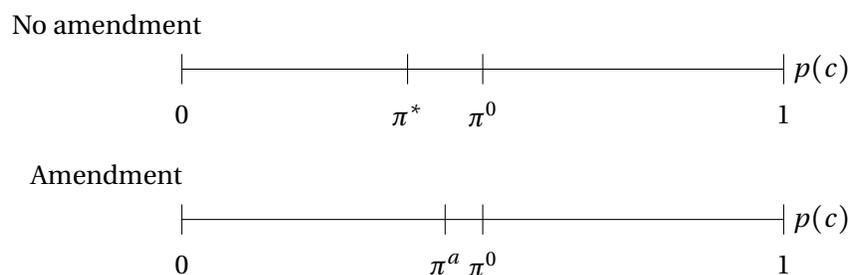


Figure 3. Threshold change for productive type with loan

set to \hat{r} . This yields a new seize threshold, π^a for the autocrat when he receives a loan.²² However, it turns out that π^a is greater than π^* , the autocrat's loan threshold without the amendment. Again, an unproductive autocrat's coup space is unchanged. But even if the lender lends, we see that the amendment decreases a productive autocrat's coup space. This is depicted in fig. 3.

Thus, we see that the amendment leaves the coup space of some autocrats unaltered — viz., unproductive challengers — while it reduces the coup space of at least some types of autocrats — viz., those autocrats who profit from receiving a loan.

5. DISCUSSION

What can we learn about Pogge's proposal from the model? More generally, what lessons does the model yield concerning the use of formal models as a tool for testing institutional design prescriptions?

The analysis in the last section partially confirms Pogge's intuition. Under certain conditions, his proposed *Odious Debt Amendment* would reduce the threat of a coup from autocrats who receive net benefit from obtaining a loan. But close analysis of the model shows that we should hesitate to endorse Pogge's proposal. First, the outcome depicted in figs. 2 and 3 depends upon stringent best-case assumptions. Central among these are that the autocrat poses no special lending risk without the amendment, that the autocrat repays his loans, and that the amendment is unproblematically enforced. Intuitively, if the autocrat posed an additional lending risk in the absence of the amendment, then the minimum acceptable rate at which the lender would be willing to lend to the autocrat

²² This is formally defined in the appendix.

would increase, perhaps as high as $\hat{r} = \frac{1-p^*(1-\gamma)}{p^*(1-\gamma)}$. If this were the case, then the amendment would leave the coup space of all types of autocrats unchanged. If the autocrat had the option of defaulting on loans once he received them, he'd be able to increase either the net benefit of holding office or the probability of staying in office by increasing the amount he spends on maintaining support (or both). The effect of problematizing enforcement is straightforward. If the amendment is not credibly enforced, then the amendment fails to affect the lender's lending decision, in which case, the autocrat's coup space remains unchanged.

The second reason we should hesitate to endorse the proposed amendment illuminates a central advantage of modelling design prescriptions formally. By formalizing our premises, we're able to keep better track of a larger set of the proposed amendment's implications. Pogge's (and our) intuition (partially) tracks the effect of the amendment on the central issue: the number of coups undertaken against fledgling democracies. But intuition is incapable of keeping track of the numerous 'peripheral' consequences, many of which are simply unanticipated. A formal model can act as a 'bookkeeping device' that enables us to examine these unanticipated consequences. I only catalogue the unanticipated consequences of the proposal. Much more detailed discussion and comparison with empirical data is warranted to make the results more compelling. Alas, such a discussion is beyond the scope of this paper.

To frame this brief discussion, consider what makes an autocrat productive or unproductive. The revenue functions track the autocrat's spending decisions. An autocrat is productive when he's able to garner a return on his income and unproductive otherwise. In general, autocrats are unproductive when their spending focuses on providing private goods for political insiders — patronage, bribery, personal aggrandizement, politically-motivated white elephant projects, etc. As paradigm examples of unproductive autocrats, consider Mobutu Sese Seko (Zaire, 1965–1997) or Ferdinand Marcos (Philippines, 1965–1986), notorious kleptocrats who openly engaged in patronage politics. In contrast, autocrats are productive when their spending focuses on investment in public goods — infrastructure, human capital accumulation (e.g., education and health), and sectoral diversification. Indonesia's Suharto (1967–1998) and China's growth since Deng Xiaoping are instructive examples here. (Of course, these examples are not void of unproductive spending practices.)²³

In light of this intuition, consider again the results of the model. The amendment deters productive autocrats. But it isn't this simple. The amendment works by shifting

²³ On the importance of private vs. public goods, see [Bueno de Mesquita et al. \(2003\)](#) and [Clark et al. \(2010\)](#).

the lower bound on the productive autocrat's coup space upward. But this only deters a subset of the productive autocrats, viz., the least stable among them $[\pi^* \leq p(c^*) < \pi^0]$. The most stable productive autocrats $[\pi^0 \leq p(c^0) < p(c^*)]$ are undeterred. However, by removing the option of receiving a loan, the amendment does decrease the stability of the undeterred productive autocrats (see lemma 5 in the appendix).

The amendment doesn't deter the worst kind of autocrat — the unproductive autocrats, the Mobutus and Marcoses. In fact, by removing the possibility of receiving loans, the amendment increases the stability of unproductive autocrats, thereby making them more difficult to remove (see lemma 3 in the appendix).

More generally, the model enables us to rigorously investigate counterfactual claims about the outcome of the amendment and keep track of a wider range of implications than we might otherwise be able to keep track of. It does this by explicitly stating premises concerning the identities of the relevant agents and the structure of their interactions (i.e., the timeline), premises concerning the agents' choices and their available options, and formalizing premises concerning the motivation and desires of the relevant agents make decisions (i.e., the utility functions). The model also enables us to isolate the causal effect of the amendment on the outcome of interest, the incidence of coups, by making certain simplifying assumptions: that the decision to enact the amendment is not endogenous to the modelled interaction; that the autocrat is successful if he undertakes a coup; that the autocrat poses no special lending risk; that the autocrat repays his loans if he remains in office; that the revenue is instantaneously realized; that the lender is the autocrat's only source of credit.

The model is certainly not a conclusive test of Pogge's proposal. There can be no conclusive test of a counterfactual claim. But the model does constitute *a* test. The model enables us to determine whether the anticipated outcome of the prescribed institutional design is a strategically logical implication of our assumptions about the relevant actors, the structure of their interactions, and their motivations and preferences. This is an important test because our intuitions about the outcomes that would follow from a particular institutional design are unreliable; intuitively plausible outcomes often turn out to be just that.

6. APPENDIX

6.1. *The Model*

To start with, I give a formal treatment of the player's payoffs. Throughout, I refer to the autocrat as *A* and the lender as *L*.

A 's objective function, once in office, is given in (1).

$$\begin{aligned} U_A(c, \lambda) &= p(c) [R(y) + V(\lambda) - c - \lambda(1+r)] + [1 - p(c)](-k) \\ &= p(c) [R(y) + V(\lambda) - c - \lambda(1+r) + k] - k \end{aligned} \quad (1)$$

Since A receives a payoff of 0 if he refrains from seizing power, A attempts a coup if and only if $U_A(c, \lambda) \geq 0$.

$R(y) + V(\lambda)$ is A 's total income while in office. $R(y) > 0$ is the government revenue generated from non-loan income sources, y , such as resource extraction. For notational simplicity, $R(y) = R$ hereafter. $V(\lambda)$ is the credit revenue generated from the loan income. I assume that $V(\lambda)$ is instantaneously realized. (For symmetry, I assume that $R(y)$ is also instantaneously realized.) $V'(\lambda) > 0$ and $V''(\lambda) < 0$. For simplicity, $\lambda = 1$ if A receives any loans and $\lambda = 0$ if A receives no loans. If A receives a loan, it must repay the lender the amount of the loan plus interest $(1+r)$, with $r > 0$. $V(1)$ can be less than, equal to, or greater than $1+r$. If $V(1) \leq 1+r$, I call A 'unproductive'. If $V(1) > 1+r$, I call A 'productive'. For notational simplicity, $V(0) = 0$ and $V(1) = V$ hereafter. $k \geq 0$ is the fixed cost to A of being removed from office.

$p(c)$ is A 's political survival probability as a function of the amount $c > 0$ that A spends on maintaining political support. c is subject to the budget constraint $c \leq M = R + V(\lambda) - \lambda(1+r)$. I assume that $p'(c) > 0$ and $p''(c) < 0$. I also assume that $p(0) = 0$ and $p(M) < 1$. Thus, $p(c) \in [0, 1)$. Finally, $\text{argmax}_c p(c) = M$.

I assume A spends the optimal amount once in office. There are two distinct optimal spending levels. If A obtains a loan ($\lambda = 1$), I denote the optimal spending level c^* and define it as follows:

$$c^* \equiv \text{argmax}_c p(c) (R + V - c - 1 - r + k) - k, \quad (2)$$

which means that

$$p'(c^*) (R + V - c^* - 1 - r + k) - p(c^*) = 0. \quad (3)$$

If A does not obtain a loan ($\lambda = 0$), I denote the optimal spending level c^0 and define it as follows:

$$c^0 \equiv \text{argmax}_c p(c) (R - c + k) - k, \quad (4)$$

which means that

$$p'(c^0) (R - c^0 + k) - p(c^0) = 0. \quad (5)$$

The optimal c is fixed by (2) or (4). Once the optimal spending level is fixed, this fixes

$p(c)$. For notational simplicity, $p(c^*) = p^*$ and $p(c^0) = p^0$ hereafter.

Claim. If $k < \frac{p(M)}{p'(M)}$ then $c^*, c^0 \in (0, M)$.

Proof. Assume that $p'(M)k - p(M) < 0$, which implies that $k < \frac{p(M)}{p'(M)}$. We can see that $c^*, c^0 \in (0, M)$ by examining (3) and (5). First, $R + V(\lambda) - c - \lambda(1+r) + k$ is positive and at its maximum when $c = 0$ and $p(0) = 0$. Second, both $p'(c)$ and $R + V(\lambda) - c - \lambda(1+r) + k$ are continuous and monotonically decreasing in c , while $p(c)$ is continuous and monotonically increasing in c . Third, these points, along with the assumption that $p'(M)k - p(M) < 0$, imply that both (3) and (5) are greater than 0 when evaluated at 0 (or at an arbitrarily small $\varepsilon > 0$ if we assume $p'(0)$ is undefined) and both are less than 0 when evaluated at M . It follows from these three points that $U'_A(c)$ is monotonically decreasing and that $U'_A(c) > 0$ for $c \in [0, \hat{c})$ and $U'_A(c) < 0$ for $c \in (\hat{c}, M]$, where \hat{c} is such that $U'_A(\hat{c}, \lambda) = 0$. From this it follows that $U''_A(c) < 0$. These conditions are sufficient to guarantee that $c^*, c^0 \in (0, M)$. \square

I assume that the lender's decision to lend or not is motivated solely by expected profit. L 's payoffs for lending are formalized in (6).

$$U_L(\lambda = 1, c) = \begin{cases} (1-\gamma)(1+r) + \gamma 0 & \text{if no amendment} \\ p(c)[(1-\gamma)(1+r) + \gamma 0] + [1-p(c)]0 & \text{if amendment} \end{cases} \quad (6)$$

L 's payoff for not lending is

$$U_L(\lambda = 0, c) = 1 \quad (7)$$

r and $p(c)$ are defined as above. I assume L is a 'price-taker' and that competition drives the market interest rate down to the minimum acceptable rate for all lenders given γ , which is set to $\underline{r} = \frac{\gamma}{1-\gamma}$.

6.2. No Amendment

Lemma 1. Given \underline{r} , L always lends to C .

This follows from the fact that $U_L(\text{lend}) \geq U_L(\text{no lend})$ when $r \geq \frac{\gamma}{1-\gamma}$.

Proposition 1. If A receives a loan, then: A seizes power iff $p^* \geq \pi^* = \frac{k}{R+V-c^*-1-r+k}$.

Proof. Assume A requests a loan. Then A spends c^* once in office and remains in office

with probability p^* . The threshold is identified by the cutpoint on the unit interval where A is indifferent between seizing power and refraining.

$$\begin{aligned} U_A(c^*, 1) &= U_A(\text{refrain}) \\ p^* (R + V - c^* - 1 - r + k) - k &= 0 \\ p^* &= \frac{k}{R + V - c^* - 1 - r + k} \end{aligned}$$

To avoid notational confusion between the probability p^* and the threshold, define the threshold as

$$\pi^* \equiv \frac{k}{R + V - c^* - 1 - r + k}. \quad (8)$$

Since A seizes power iff $U_A(c^*, 1) \geq U_A(\text{refrain})$, A seizes power iff $p^* \geq \pi^*$. \square

Proposition 2. *If A does not receive a loan, then: A seizes power iff $p^0 \geq \pi^0 = \frac{k}{R - c^0 + k}$.*

Proof. Assume A does not receive a loan. Then A spends c^0 once in office and remains in office with probability p^0 . The threshold is defined as the cutpoint on the unit interval where A is indifferent between seizing power and refraining.

$$\begin{aligned} U_A(c^0, 0) &= U_A(\text{refrain}) \\ p^0 (R - c^0 + k) - k &= 0 \\ p^0 &= \frac{k}{R - c^0 + k} \end{aligned}$$

To avoid notational confusion between the probability p^0 and the threshold, define the threshold as

$$\pi^0 \equiv \frac{k}{R - c^0 + k}. \quad (9)$$

Since A seizes power iff $U_A(c^0, 0) \geq U_A(\text{refrain})$, A seizes power iff $p^0 \geq \pi^0$. \square

Before proving the next lemmas, define $F(c)$ and $G(c)$ as follows.

$$\begin{aligned} F(c) &\equiv p'(c)(R - c + k) - p(c) \\ G(c) &\equiv p'(c)(R + V - c - 1 - r + k) - p(c) \end{aligned}$$

Comparing $F(c)$ and $G(c)$, we see that

$$G(c) = F(c) + p'(c)(V - 1 - r).$$

From (3) and (5), we know that $F(c^0) = 0$ and $G(c^*) = 0$.

Lemma 2. *If $V \leq 1 + r$ then $c^0 \geq c^*$.*

Proof. Suppose $V = 1 + r$. Then $p'(c^0)(V - 1 - r) = 0$ and $G(c^0) = 0$. It follows that $c^0 = c^*$.

Suppose $V < 1 + r$. Since $p'(c) > 0$, $p'(c^0)(V - 1 - r) < 0$. Thus, $G(c^0) < 0$. Since c^* is an interior maximum, it follows that c^0 is to the right of c^* , which means that $c^0 > c^*$.

Thus, if $V \leq 1 + r$, $c^0 \geq c^*$. \square

Lemma 3. *If C is unproductive, then C is at least as likely to remain in office without a loan as with a loan ($p^0 \geq p^*$).*

This follows from lemma 2 and the fact that $p(c)$ is monotonically increasing in c .

Lemma 4. *If $V > 1 + r$ then $c^0 < c^*$.*

Proof. Suppose $V > 1 + r$. Since $p'(c) > 0$, $p'(c^0)(V - 1 - r) > 0$. Thus, $G(c^0) > 0$. Since c^* is an interior maximum, it follows that c^0 is to the left of c^* , which means that $c^0 < c^*$.

\square

Lemma 5. *If C is productive, then C is less likely to remain in office without a loan than with a loan ($p^0 < p^*$).*

Proposition 3. *If $V \leq 1 + r$ then $\pi^0 \leq \pi^*$.*

Proof. Suppose $V = 1 + r$.

$$\begin{aligned} U_A(c^*, 1) &= p(c^*)[R - c^* + k + V - 1 - r] - k \\ &= p(c^*)[R - c^* + k] - k \end{aligned}$$

Since $c^* = c^0$ (from lemma 2), $U_A(c^*, 1) = U_A(c^0, 0)$ [from (1)], which means that A is indifferent between seizing power with a loan and seizing power without a loan. It follows that $\pi^* = \pi^0$.

Now suppose $V < 1 + r$. Recall the definitions of π^0 and π^* [given above in (8) and (9)]. $\pi^0 < \pi^*$ iff $c^0 - c^* < 1 + r - V$.

$$\begin{aligned} \pi^0 < \pi^* \\ \frac{k}{R - c^0 + k} < \frac{k}{R + V - c^* - 1 - r + k} \\ c^0 - c^* < 1 + r - V \end{aligned} \quad (10)$$

From (3), it follows that $c^* = R + V - 1 - r + k - \frac{p^*}{p'(c^*)}$. From (5), we get $c^0 = R + k - \frac{p^0}{p'(c^0)}$. Substituting into (10), we get

$$\begin{aligned} R + k - \frac{p^0}{p'(c^0)} - \left[R + V - 1 - r + k - \frac{p^*}{p'(c^*)} \right] < 1 + r - V \\ \frac{p^*}{p'(c^*)} < \frac{p^0}{p'(c^0)} \end{aligned} \quad (11)$$

Given that $V < 1 + r$, it follows from lemma 3 that $p^0 > p^*$. From the concavity of $p(\cdot)$, it follows that $p'(c^0) < p'(c^*)$. Thus, (11) holds. Consequently, (10) holds, from which it follows that $\pi^0 < \pi^*$.

Thus, if $V \leq 1 + r$, $\pi^0 \leq \pi^*$. \square

Proposition 4. *If $V > 1 + r$ then $\pi^0 > \pi^*$.*

Proof. The proof follows the same reasoning as in the case when $V < 1 + r$. The difference is that I must now show that $\frac{p^*}{p'(c^*)} > \frac{p^0}{p'(c^0)}$, which follows from lemma 5 and the concavity of $p(\cdot)$. \square

6.3. Amendment

Lemma 6. *Given the amendment, γ , and p^* , L lends to C if and only if $r \geq \hat{r} = \frac{1 - p^*(1 - \gamma)}{p^*(1 - \gamma)}$.*

This follows from the fact that L lends if and only if doing so yields a greater payoff than not lending. This is true when $r \geq \frac{1 - p^*(1 - \gamma)}{p^*(1 - \gamma)}$.

Proposition 5.

- (1) *If $r < \frac{1 - p^*(1 - \gamma)}{p^*(1 - \gamma)}$, then: (a) When $V < 1 + r$, A 's coup space is unchanged; (b) When $V > 1 + r$, A 's coup space is reduced to $p^0 \geq \pi^0$.*

- (2) If $r \geq \frac{1-p^*(1-\gamma)}{p^*(1-\gamma)}$, then: (a) When $V < 1+r$, A 's coup space is unchanged; (b) When $V > 1+r$, A 's coup space is reduced to $p^* \geq \pi^a > \pi^*$.

Proof. Case 1. Suppose $r < \frac{1-p^*(1-\gamma)}{p^*(1-\gamma)}$. Accordingly, the effect of the amendment is to remove the option of seizing power and receiving a loan. Now that loans are no longer available, π^* is no longer relevant. Thus, the lower bound of A 's coup space is now defined by the location of π^0 . If A is unproductive, it follows from proposition 3 that the size of the coup space is unchanged. If A is productive, it follows from proposition 4 that the amendment reduces A 's coup space.

Case 2. Suppose $r = \hat{r}$. If A is unproductive, it follows from proposition 3 that the size of the coup space is unchanged.

If A is productive, it follows from prop. 4 that the lower bound of A 's coup space is defined by π^* . To see the effect of the amendment on π^* when A is productive, substitute $\hat{r} = \frac{1-p^*(1-\gamma)}{p^*(1-\gamma)}$ into (8). Solving for p^* , we see that A 's seize threshold when it receives a loan is now

$$\pi^a \equiv \frac{1+k(1-\gamma)}{(R+V-c^*+k)(1-\gamma)}. \quad (12)$$

Now notice that $\underline{r} < \hat{r}$ for all $p^* < 1$.

$$\begin{aligned} \frac{\gamma}{1-\gamma} &< \frac{1-p^*(1-\gamma)}{p^*(1-\gamma)} \\ p^*\gamma &< 1-p^*+p^*\gamma \\ p^* &< 1 \end{aligned}$$

Since

$$\frac{\partial \pi^*}{\partial r} = \frac{k}{(R+V-c^*-1-r+k)^2} > 0 \quad (13)$$

and $\underline{r} < \hat{r}$ for all $p^* < 1$, it follows that $\pi^* < \pi^a$. Thus, even if $r \geq \frac{1-p^*(1-\gamma)}{p^*(1-\gamma)}$, we can see that the amendment decreases a productive A 's coup space. \square

REFERENCES

- AKERLOF, GEORGE A. 1970. 'The Market for "Lemons": Quality Uncertainty and the Market Mechanism'. *Quarterly Journal of Economics* 84: 488–500.
- BRENNAN, GEOFFREY and PHILIP PETTIT. 2005. 'The Feasibility Issue'. In *Oxford Handbook of Contemporary Philosophy*. Ed. by Frank Jackson and Michael Smith. Oxford: Oxford University Press.
- BUCHANAN, ALLEN. 2004. *Justice, Legitimacy, and Self-Determination: Moral Foundations for International Law*. New York: Oxford University Press.
- BUENO DE MESQUITA, BRUCE et al. 2003. *The Logic of Political Survival*. Cambridge, MA: MIT Press.
- CLARK, WILLIAM R. et al. 2010. 'Why Do Autocrats Overachieve? Political Competition and Material Well-Being in Comparative Perspective'. Unpublished manuscript.
- COLLIER, PAUL and ANKE HOEFFLER. 1998. 'On Economic Causes of Civil War'. *Oxford Economic Papers* 50: 563–573.
- 2004. 'Greed and Grievance in Civil War'. *Oxford Economic Papers* 56 (4): 563–595.
- COWEN, TYLER. 2007. 'The Importance of Defining the Feasible Set'. *Economics and Philosophy* 23 (1): 1–14.
- ELSTER, JON. 2007. *Explaining Social Behavior: More Nuts and Bolts for the Social Sciences*. New York: Cambridge University Press.
- FEARON, JAMES D. and DAVID D. LAITIN. 1996. 'Explaining Interethnic Cooperation'. *American Political Science Review* 90 (4): 715–735.
- HEDSTRÖM, PETER and RICHARD SWEDBERG, eds. 1998. *Social Mechanisms: An Analytical Approach to Social Theory*. New York: Cambridge University Press.
- JENSEN, MARK. 2009. 'The Limits of Practical Possibility'. *Journal of Political Philosophy* 17 (2): 168–184.
- KNUUTTILA, TARJA. 2009. 'Isolating Representations Versus Credible Constructions? Economic Modelling in Theory and Practice'. *Erkenntnis* 70 (1): 59–80.
- KUORIKOSKI, JAAKKO and AKI LEHTINEN. 2009. 'Incredible Worlds, Credible Results'. *Erkenntnis* 70 (1): 119–131.
- MÄKI, USKALI. 2005. 'Models are Experiments, Experiments are Models'. *Journal of Economic Methodology* 12 (2): 303–315.
- 2009. 'MISSing the World. Models as Isolations and Credible Surrogate Systems'. *Erkenntnis* 70 (1): 29–43.
- MELIA, JOSEPH. 2003. *Modality*. Montréal: McGill-Queen's University Press.
- POGGE, THOMAS W. 2002. *World Poverty and Human Rights*. Malden, MA: Polity Press.

- RAIKKA, JUHA. 1998. 'The Feasibility Condition in Political Theory'. *Journal of Political Philosophy* 6 (1): 27–40.
- REDDY, SANJAY. 2005. 'The Role of Apparent Constraints in Normative Reasoning: A Methodological Statement and Application to Global Justice'. *The Journal of Ethics* 9: 119–125.
- SCHELLING, THOMAS C. 1978. *Micromotives and Macrobehavior*. London and New York: W. W. Norton & Co., 2006.
- WENAR, LEIF. 2008. 'Property Rights and the Resource Curse'. *Philosophy & Public Affairs* 36 (1): 2–32.
- WIENS, DAVID. 2010. 'Taking the Ideal Out of Nonideal Theory: Institutional Design as Failure Analysis'. Unpublished manuscript. URL: <http://www-personal.umich.edu/~wiens>.