The Politics of the Slippery Slope*

Giri Parameswaran[a], Gabriel Sekeres[b] , and Haya Goldblatt[c]

[a]Haverford College, gparames@haverford.edu
[b]Stanford Institute for Economic Policy Research, gsekeres@stanford.edu
[c]Haverford College, haya.goldblatt001@gmail.com

April 14, 2024

Abstract

Slippery slope arguments — the idea that otherwise beneficial reforms should be rejected lest they beget further undesirable one — are ubiquitous in political discourse. We provide a learning-based policy-feedback mechanism to explain why slippery slope dynamics arise. Additionally, we provide conditions under which, in equilibrium, sophisticated agents will successfully manipulate policy to either induce or prevent a slippery slope dynamic.

Key Words: Slippery slope, Learning, Policy momentum.

*Supplementary material for this article is available in the Appendix in the online edition. The authors did not receive any financial support during the production of this manuscript.

1 Introduction

'Slippery slope' arguments are commonly invoked in political discourse. They express the idea that even though a policy may be desirable on its own merits, it should nevertheless be rejected because of the fear that its adoption will cause more extreme (and undesirable) policies to arise in the future.[1]

The public discourse surrounding the Affordable Care Act (ACA) provides a useful case study. Despite largely mirroring a proposal from the conservative Heritage Foundation, and notwithstanding its adoption by a Republican administration in Massachusetts, the ACA did not command the support of congressional Republicans, and was even met with suspicion by conservative Democrats. For example, during negotiations over the bill, Democratic Senator Ben Nelson expressed opposition to a proposed Medicare buy-in worrying that it would be a "forerunner of single payer" healthcare (Raju 2009). His concern was not unfounded. After the ACA had passed, then Democratic Senate Majority Leader Harry Reid confirmed that his goal was "absolutely" to transition the ACA to a single payer system (Roy 2013). The consideration by courts of the ACA's legality also raised slippery slope concerns. Justice Antonin Scalia famously worried that, absent a clear limiting principle, a government mandate to buy health insurance today would invite future governments to mandate the purchase of more mundane items such as broccoli.

Both examples have the feature that the immediate policy in question acts as a stepping stone that makes possible a more extreme policy, which would be politically infeasible to implement directly today, but which might become feasible as the public becomes accustomed to the moderate change. The slippery slope dynamic is generated by policy feedback: experience with a moderate policy may cause the public to re-evaluate their beliefs about the value of that reform, and potentially demand even more of it. This insight reflects Schattschneider's (1935) aphorism that 'a new policy creates a new politics'.

1. For examples of slippery slope arguments, see Dent (1999) and Kurtz (2003) on same-sex marriage, Nix (2012) and Somin (2012) on the Affordable Care Act mandate, and Volokh (2003) for an exhaustive primer.

Implicit in the logic of policy feedback is that agents learn about the value of certain policies as they interact, and become acquainted with, those (or similar) policies. Experience with ACA programs, for example, has been shown to positively influence agents' opinions about the ACA, as well as other governmental healthcare schemes, such as Medicare (see Lerman and McCabe 2017; Jacobs and Mettler 2018; Campbell 2020). This policy feedback occurs in many other contexts, as well (see below). Importantly, the literature finds that the extent of this feedback depends on a range of factors, particularly the size, scope, and import of the policy, and the likelihood that agents experience or engage with it directly.

In this paper, we first explore a particular mechanism that explains why a slippery-slope dynamic — in which a moderate reform today begets a more extreme reform in the future — might arise. We also investigate the conditions under which (some) agents' awareness of this policy feedback might create an incentive to strategically manipulate policy to either induce or prevent the dynamic from arising.

To answer these questions, we present a simple stylized model of public goods provision under majority rule. Agents are distinguished by their income; a majority have low income whilst the remainder have high income. Low income earners have a higher demand for the public good than high income earners, and this generates the baseline political disagreement between the groups.

Additionally, each agent may either be correctly informed about the value of the public good, or misinformed. We focus most attention on the case where misinformed agents *undervalue* the public good; thus expressing a lower demand than their informed counter-parts. This reflects the public's typical skepticism towards unfamiliar projects and reforms. We discipline the model by assuming that a majority of agents are informed — so that our results are not purely driven by misinformed majorities. Importantly, though a majority are poor and a majority are informed, we assume that the informed poor are a minority.[2]

2. If they were a majority, then they would constitute a decisive coalition in their own right, and there would be no interesting political economy analysis.

Misinformation has two effects: First, the median voter's preferred level of public goods provision will be below that of the informed poor. There will be policy 'skepticism' relative to the 'correct information' baseline. Second, the preferences of the misinformed poor and the informed rich will be more closely aligned, and these groups may potentially form a cohesive voting bloc, even though their intrinsic preferences (if correctly informed) diverge.

We consider a simple learning-by-acquaintance technology wherein agents learn the correct value of the public good whenever it is provided in a sufficiently large quantity to be consequential to their utility. This is consistent with empirical findings, noted above, that learning about policy is strongest when the policy is salient and visible to the agent.

Taken together, these features of our model imply several noteworthy results. First, if learning occurs in some period, it causes the ranks of the informed to grow, which increases future social demand for the public good, *ceteris paribus*. Learning shifts political power between the different groups. A moderate policy today combined with a skeptical public who can learn from acquaintance, induces a more extreme policy tomorrow. This is the slippery-slope dynamic at work. Policy momentum arises endogenously, as a consequence of learning by acquaintance.

Second, since the slippery slope dynamic hurts the informed rich (by moving policy farther from their ideal), they have an incentive to downwardly distort policy to prevent learning. To be successful, the informed rich must enlist the support of the misinformed poor, to build a majority coalition around this distorted policy. But this can only occur if the ideal policy of the misinformed poor is *even lower* than that of the informed rich (absent strategic considerations). Thus, strategic manipulation of policy will only occur if misinformation creates a larger wedge in policy preferences between the informed and misinformed poor than is the inherent wedge between the (informed) rich and poor, ensuring that a natural alliance exists between the informed rich and misinformed poor against the informed poor.

Of course, distorting policy is costly to the informed rich, and so the incentive to behave

strategically extends only as far as the benefits from preventing learning exceed the costs. This requires that the distortion needed to prevent learning is not too large.

The logic of strategic behavior requires that some agents are forward looking, and understand the policy dynamic that arises when there is learning by acquaintance. Our third result suggests a complement to this insight: the ability of some agents to strategically manipulate policy is limited by the degree of sophistication of other agents, and their awareness of being manipulated. The informed rich will be most able to strategically prevent the slippery slope dynamic when sufficiently many misinformed agents are myopic. By contrast, if the misinformed poor are sophisticated in sufficient numbers, then opportunities for strategic manipulation will disappear. Moreover, the competing incentives to manipulate and prevent manipulation will often result in policy incoherence, where no equilibrium policy exists.

In light of the motivating example, the public finance setting is a natural one to study the interplay between beliefs, intrinsic preferences (captured by income), and policy. However, the model's insights can be extended beyond this narrow setting. What is important is the interaction between the nature of misinformation and the learning technology, such that those who seek to strategically prevent learning can make common cause with the misinformed.

As a counter-point, we note (in Appendix A.2) that policy neither evolves endogenously nor is there policy distortion if the misinformed minority *overvalue* the public good, and are thus *optimistic* (rather than skeptical) of the policy reform. The reason is that the political incentives now pull in opposite directions; the informed rich still prefer a lower policy than the informed poor, but the misinformed poor will now prefer a higher policy than the informed rich. A natural coalition that supports strategic under-provision of the public good no longer exists.[3]

3. Our maintained assumption that a majority of agents are informed plays an important role here. If instead, a majority are optimistic, then there can be an equilibrium with policy reversal; a larger policy is enacted in the first period that begets learning, and a subsequent roll-back of policy in the second period. The 'war on terror' and Brexit are both policies that majorities initially supported but came to later regret.

Somewhat more subtly, we show that in some instances, misinformation can cause policy to be over-provided. In particular, a pessimistic minority (who perceive the majority to be overly optimistic) may have an incentive to strategically over-provide the public good in order to induce learning that teaches the majority that they have been too optimistic. A similar dynamic can arise when the misinformed are a majority and optimistic. Here, learning serves a 'teach them a lesson by giving them what they want' flavor. Our analysis highlights conditions under which such a dynamic may arise.

Though perhaps less commonplace, examples of the lesson-teaching motive exist. Consider the response from law enforcement to the 'defund the police' movement. In many instances, police responded by withdrawing services, for example by patrolling neighborhoods less intensely. A widely reported case was the occupation of the Capital Hill area of Seattle in 2020 as part of broader anti-policy violence protests. The Seattle Police Department responded by abandoning the area, enabling the creation of the Capital Hill Autonomous Zone. The result, as examined by Piza and Connealy (2022), was an increase in crime in the zone and the surrounding area of Seattle. The CHAZ ended just three weeks after it began, and the police re-occupied the area. In over-providing the policy reform of 'less police', law enforcement were seemingly able to demonstrate that the benefits of such a policy were overstated.[4] More generally, following short periods of reduced funding for police, many communities have restored or even increased that funding (see Fegley and Murtazashvili 2023).

The two key behavioral assumptions under-pinning our model — that there is policy feedback and that agents often undervalue reforms — are supported by considerable empirical evidence. First, there is strong evidence that policy feedback occurs, and that voters learn by acquaintance. For example, Campbell (2011) finds that the introduction of Social Security led to both increased knowledge of the program and increased support for it. Similarly, Cook,

4. For clarity, in this example, we take the policy reform in question to be a movement away from intensive policing. Of course, police services are quite straightforwardly public goods, and in other contexts, 'over-provision' would involve hiring too many police. Instead, in this context, the reform precisely 'does more' by policing less (coupled with mental health and other community-oriented types of outreach).

Jacobs, and Kim (2010) find that sending information on social security to a random sample of beneficiaries increased confidence in the program among those who received the information. In a different context, Baccini and Leemann (2012) show that voters are more likely to be sensitive to climate issues after being exposed to a natural disaster. Similar effects exist regarding attitudes towards gay and lesbian people. Herek and Glunt (1993) and Herek and Capitanio (1996) show that interpersonal contact was the strongest predictor of positive attitudes towards homosexuals. And, public policy affects the opportunities for learning by acquaintance to occur. Day and Schoenrade (1997, 2000) and Griffith and Hebl (2002) show that anti-discrimination policies cause individuals to be more open about their sexuality, thereby enabling known interpersonal contact between homosexuals and heterosexuals.

Second, the idea that voters are often mistaken about the value of reforms or public goods is also plainly evident. In a survey study, Koch and Mettler (2012) found that over 50% of respondents receiving some type of government benefit (such as the Home Mortgage Interest Deduction, the Earned Income Tax Credit, Pell Grants or Food Stamps) were unaware that those benefits were indeed provided by the government. This perceived absence of government in their lives suggests that agents will be more skeptical of the value of public spending than they would ideally, if they were correctly informed. Conversely, when government spending is seen to be wasteful or directed towards ends that do not directly improve the public welfare, voters tend to inflate the costs of such programs. U.S. spending on foreign aid provides a stark example. The median respondent in a 2010 World Public Opinion Poll of 848 Americans believed that the foreign aid budget accounted for 25% of the federal spending, whilst only 19% believed it was below 5%. In fact, it was less than 1% of total federal spending. By over-attributing the share of public spending on 'non-beneficial' projects, voters effectively undervalue public spending as an aggregate bundle.

Moreover, history is replete with examples of policies that voters were originally suspicious or skeptical about, but eventually came to appreciate. Social Security, which is now extremely

popular amongst voters, was, at its inception, feared by many as a socialist scourge that would enslave Americans.[5]

This paper contributes to, and extends, several strands of the political economy literature. At its core, the inefficiency in this model arises from an endogenous time inconsistency in the decision makers' preferences, arising out of the changing identity of the pivotal voter. This feature is common to many models of inefficiencies in policy making, including Persson and Svensson (1989), Roberts (1989), Alesina and Tabellini (1990), Dewatripont and Roland (1992), Benabou (2000), and Battaglini and Coate (2008), amongst many others. However, in contrast to many of those models, and similar to Acemoglu and Robinson (2001), Benabou and Ok (2001), the shifting political power is not exogenous, but endogenous to the current pivotal agent's policy choice, in our model.

Policy momentum is another feature of this model that is present in Benabou and Ok (2001). In that paper, policy is sticky. This creates a fear in the current poor that a redistributive policy that will benefit them in the short run will persist long enough to eventually expropriate their future wealth. Thus, policy inertia is hard-wired into their model. Our paper is more standard in that it allows the polity to change its policy in every period. Reform momentum arises as an equilibrium phenomenon rather than as a feature of the model technology.

Besley and Coate (1998) present a model, similar to ours, with endogenously evolving policy. In the first period, the polity must both choose a (redistributive) tax policy and decide whether to undertake a public investment that changes the distribution of second period incomes. In equilibrium, the public investment may be rejected even when it weakly increases all incomes, if the change in the income distribution causes political power to shift. A similar dynamic exists in Bénabou, Ticchi, and Vindigni (2022), where a first period investment (in technologies that suppress scientific discoveries) affects second period preferences for religious

5. Unsurprisingly, slippery slope concerns formed part of the objection to Social Security. During Congressional hearings, a senator from Oklahoma asked Secretary of Labor Frances Perkins, "Isn't this socialism?". When she answered no, he responded: "Isn't this a teeny-weeny bit of socialism?" Altman (2005)

public goods.

Though both these papers involve policy feedback, we think that our approach, which highlights the feedback between policy and beliefs, is apt for the slippery slope context. In our model, it is the first period tax policy *itself* that affects preferences over the second period tax policy — and this makes the connection to policy momentum natural. By contrast, in both Besley and Coate (1998) and Bénabou, Ticchi, and Vindigni (2022), policy preferences and beliefs are affected through some separate dimension of first period policy (e.g. public investment). For example, in Besley and Coate (1998), high first period redistribution doesn't itself impact the second period preference of any voter — it is the additional choice of whether to engage in public investment or not that does so.

A related literature examines the incentives for policymakers to distort (or even sabotage) policy. Groseclose and McCarty (2001) and Hirsch and Kastellec (2022) both present models in which policy distortion serves to harm the reputation of an incumbent whose ability is imperfectly observed by voters. Kang (2022) explores the opposite logic of a Congress that is overly deferential to the president, in the hope of showcasing subsequent presidential failure — analogous to the lesson-teaching motive for over-providing the public good in our model.

Finally, this paper extends upon a growing literature on learning in a political economy context. Fernandez and Rodrik (1991) consider a model in which asymmetric information about the identity of winners and losers from a reform may cause the reform to fail, even if the reform makes the average agent better off. Similar to this paper, they find an endogenous bias towards status quo policies. More recent work consider the incentives for agents to choose policies that affect the learning of others. Strulovici (2010) studies learning in bandit problems where decisions (about how to experiment) are made collectively by majority vote. Heidhues, Koszegi, and Strack (2018) develop a model of individual Bayesian learning with an overconfident prior. Hirsch (2014) studies the dynamic interaction between a principal and agent who share common preferences but different beliefs, where learning is possible from

past choices. Baker and Mezzetti (2012), Fox and Vanberg (2014), and Parameswaran (2018) consider models of the judiciary in which learning occurs after courts observe the outcomes of agent choices. In the electoral setting, Dewan and Hortala-Vallve (2019) present a model voters learn about an incumbent's ability based on the success of past reforms. In each case, the learning motive distorts the incentives to efficiently provide the reform.

The rest of the paper is organized as follows. Section 2 introduces the characteristics of the model. Section 3 establishes basic analytical insights, and Section 4 analyzes the model in a dynamic equilibrium setting. Section 5 concludes. All proofs, as well as several additional extensions, appear in the Appendix.

2 Model

We present a dynamic model with two periods, $t = 1, 2$. There is a unit mass of agents. Each agent may either be rich or poor. Poor agents have (exogenous) income $y_L > 0$ while rich agents have income $y_H > y_L$.[6] We assume that a majority of agents are poor, so that the median income earner has low income.

In each period, the government can provide a quantity $g \geq 0$ of a public good. The public good has unit cost normalized to 1, and is financed through a non-distortionary, proportional tax on income, $\tau \in [0, 1]$. The government's budget constraint is $g = \tau \overline{y}$, where $\overline{y}$ is the average per-capita income.

An agent with income y_i has preferences over feasible policies (τ, g) given by

$$u(\tau, g; y_i) = (1 - \tau)y_i + A \ln g$$

where A parameterizes the marginal benefit of public good spending. The log-linear functional form choice is purely to keep expressions simple; the basic insights will continue to hold for any concave preference.

6. In Appendix A.4 we explore an extension where agent incomes are drawn from a continuous distribution. All the key insights continue to hold in that richer setting.

Each agent may either be correctly informed (I) or misinformed (M) about the value of the public good. Informed agents know the true value of A (which we denote by A_I), while misinformed agents believe that it takes a different value A_M. In our main analysis, we assume that misinformed agents undervalue the public good ($A_M < A_I$), as this will be shown to be the most interesting case. In Appendix A.2 we consider the opposite case of misinformed agents who overvalue the good ($A_M > A_I$). To ensure that the first order conditions produce interior solutions, we assume that $A_I < y_L$.

So far, we have identified four types of agents; each agent having one of two possible incomes and one of two possible beliefs. For each type $i \in \{LI, HI, LM, HM\}$, let ϕ_i denote the proportion of type-i agents in the economy. No group constitutes a majority in its own right, so $\phi_i < \frac{1}{2}$ for all i. As above, we assume that a majority of agents are poor (i.e. $\phi_{LI} + \phi_{LM} > \frac{1}{2}$). To ensure that our results are not purely driven by the existence of a large number of misinformed voters, we assume that a majority of agents are informed (i.e. $\phi_{LI} + \phi_{HI} > \frac{1}{2}$). Finally, for technical convenience, we assume that the informed rich and the misinformed poor together constitute a majority (i.e. $\phi_{HI} + \phi_{LM} > \frac{1}{2}$). This latter assumption simplifies the analysis, though our insights will continue to hold even if the condition is violated. Taken together, these assumptions imply that any two of the three largest groups — informed poor, informed rich, and misinformed poor — will together constitute a majority.

We study a simple and stark model of learning. In each period, after a policy is implemented, each agent compares their actual utility against the utility they were expecting, given their belief about A. When these are sufficiently different, the agent realizes that their belief must have been incorrect, and updates their belief to the true value. Formally, an agent with belief A learns whenever:

$$|u(\tau, g; y_i, A) - u(\tau, g; y_i, A_I)| > \mu$$

where $\mu > 0$ parameterizes the agent's sensitivity to information.[7] Though stark, this learning technology operationalizes our story in the simplest possible way. In Appendix A.5, we show that our insights would continue to hold if agents were Bayesian and had non-degenerate priors — though at the cost of considerable complexity.

Finally, agents may either be sophisticated or myopic. A sophisticated agent understands that learning by acquaintance in the first period affects the polity's second period beliefs (and thus policy preferences). When evaluating policies in the first period, sophisticated agents take this dynamic effect into account. Myopic agents, by contrast, ignore this dynamic effect, and so evaluate policies purely based on their stage game payoff. An alternative interpretation of sophisticated and myopic types is that all agents understand the learning dynamics, but that sophisticated agents are future oriented (putting weight $\beta > 0$ on future utility), whilst myopic agents are purely present oriented (i.e. with discount factor 0). Sophistication and myopia only have their bite in the first period. Since the game ends after the second period, neither type entertains dynamic policy considerations in the second period.

A note about sophistication is in order. Though the model identifies some agents as informed and others as misinformed, all agents will naturally perceive themselves as being correctly informed. Thus, a sophisticated agent with belief A will also believe that, whenever there is learning, other agents will come to share *their* belief.

There are two sophisticated political parties that are purely office motivated. In each period, each party announces a feasible fiscal policy (τ, g) that it is committed to implement if elected. Voters cast their ballots and the party receiving a majority of the vote is elected. A feasible policy (τ, g) is a majority winner if it is preferred to any other feasible policy (τ', g') by a majority of agents. The median voter theorem predicts that competition between the parities

7. In Appendix A.3, we explore an extension in which agents have heterogeneous sensitivities to information. We show that the key insights of our baseline model continue to hold.

will lead them both to propose majority winning policies. Thus, we associate equilibrium with the majority winning policy whenever it exists.

3 Preliminaries

3.1 Stage Game Benchmark

We begin by studying the equilibrium in a single period game absent dynamic considerations. Since the game is static, the agents evaluate policies by their associated stage game utilities.

Consider an agent with income y and whose belief about the value of the public good is A. We make the standard assumption that all agents understand the government's budget constraint; there is no fiscal illusion. Recall that the budget constraint is $g = \tau\overline{y}$; the quantity of public goods provided is in direct proportion to the tax rate. A type-(y, A) agent's indirect utility over tax policies is given by:

$$v(\tau; y, A) = u(\tau, \tau\overline{y}; y, A) = (1 - \tau)y + A\ln(\tau\overline{y})$$

It is easily verified that v is strictly concave — and therefore single peaked — in τ for each (y, A). By the first order condition, a type-(y, A)'s most preferred policy is:

$$\tau^*(y, A) = \frac{A}{y} = \frac{A_I}{y \cdot \frac{A_I}{A}} = \tau^*(x(y, A), A_I) \tag{1}$$

where $x(y, A) = y \cdot \frac{A_I}{A}$. The first equality gives a direct expression for $\tau(y, A)$ as a function of y and A. All else equal, agents who believe that public goods are more valuable will demand a higher tax rate to fund more public goods; and richer agents will demand a lower tax rate and fewer public goods than poorer agents. For notational convenience, we denote a type i's ideal stage game policy by τ_i, where $i \in \{LI, HI, LM, HM\}$.

The final equality in (1) reveals that the most preferred tax rate of a type-(y, A) agent coincides with the most preferred tax rate of a type-$(x(y, A), A_I)$ agent; i.e. an informed agent having income $x(y, A)$. We refer to $x(y, A)$ as the agent's 'effective income'. It is

the income for which their most preferred policy would be truly optimal if they had correct beliefs. Naturally, the effective income of an informed agent is simply their income. However, since misinformed agents undervalue the public good ($A_M < A_I$), their effective income will be larger than their true income (i.e. $x(y, A_M) > y$). A misinformed agent who undervalues public goods expresses identical preferences to an agent with higher income who correctly values public goods.[8] Let x_i denote the effective income of a type-i agent.

Now recall that, fixing beliefs, agents' most preferred tax rates are decreasing in incomes. Hence, if the agents are ordered by their effective incomes, their most preferred policies will be monotone in that ordering. Then, since preferences are single peaked, the median voter theorem applies. The equilibrium tax rate will be the most preferred tax rate of the agent with the median effective income.

Since $y_H > y_L$ and $A_I > A_M$, it follows that the informed poor have the lowest effective income, and the misinformed rich have the highest effective income. The ordering of the remaining types' effective incomes depends on the size of the belief disagreement relative to the size of income disparities. Suppose that the divergence in beliefs is small relative to the disparity in incomes (formally, that $\frac{A_I}{A_M} < \frac{y_H}{y_L}$). Then, $x_{LI} < x_{LM} < x_{HI} < x_{HM}$, which implies that $\tau_{HM} < \tau_{HI} < \tau_{LM} < \tau_{LI}$. Because the belief distortions are small, the ideal policies of the informed and misinformed poor will be closer together than the ideal policies of the informed poor and informed rich. Low income earners, as a group, will have a larger demand for public goods than high income earners. If so, since low income earners collectively form a majority, but the informed poor are a minority, it must be that the misinformed poor are pivotal.

By contrast, if $\frac{A_I}{A_M} > \frac{y_H}{y_L}$, then $x_{LI} < x_{HI} < x_{LM} < x_{HM}$, which implies that $\tau_{HM} < \tau_{LM} < \tau_{HI} < \tau_{LI}$. The distortion in beliefs is sufficiently large that the ideal policy of the misinformed poor is further from that of the informed poor than is the ideal policy

8. Indeed, it is not just that the agents share the same ideal policies. Their preferences coincide. To see this, note that $v(\tau; y, A) = \frac{A}{A_I}\left[(1-\tau)\cdot y\frac{A_I}{A} + A_I \ln(\tau\overline{y})\right] = \frac{A}{A_I}v(\tau; x(y, A), A_I)$.

of the informed rich. Informed agents, as a group, have larger demand for public goods than misinformed agents. Then, since informed agents collectively form a majority, but the informed poor are a minority, it must be that the informed rich are pivotal.

In general, the median effective income will be $x_{med} = \min\{x_{LM}, x_{HI}\}$. The equilibrium policy will be $\tau_{med} = \frac{A_I}{x_{med}} = \max\{\tau_{LM}, \tau_{HI}\}$, and $g_{med} = \tau_{med}\overline{y}$.

3.2 Learning

There will be learning if, given feasible policy $(\tau, \tau\overline{y})$ that is implemented, an agent's anticipated utility (given their belief of A) differs sufficiently from their realized utility (given the true A). This will be true if:

$$|[(1-\tau)y + A_I \ln(\tau\overline{y})] - [(1-\tau)y + A_M \ln(\tau\overline{y})]| > \mu$$

$$\tau > \frac{1}{\overline{y}} \exp\left\{\frac{\mu}{|A_I - A_M|}\right\} = \tau^\dagger$$

i.e. if the implemented policy is sufficiently large. This captures the ideas previously discussed, that agents typically learn from policy only if the policy is sufficiently salient. $\tau^\dagger$ denotes either the highest policy that does not result in learning, or the lowest policy that induces learning.[9] Notice that $\tau^\dagger$ is increasing in μ and decreasing in $|A_I - A_M|$. Intuitively, the less sensitive agents are to information (i.e. the higher is μ), the more salient the policy must be to induce learning. By contrast, the larger is the disparity in beliefs, the less salient the policy needs to be to convince the agent. Since there is a one-to-one relationship between μ and $\tau^\dagger$ (holding $|A_I - A_M|$ fixed), it suffices to present results in terms of $\tau^\dagger$ rather than μ. For simplicity, we will assume that $\tau^\dagger < \tau_{LI}$, which rules out the possibility of no learning even when the highest stage-game consistent policy is chosen.

Note that an agent's propensity to learn is independent of their income. In any period there will be learning by misinformed agents of both income types, or neither income type. Given

9. For technical reasons related to maximizing on an open set, we need to allow either possibility at the threshold. We could avoid the ambiguity by discretizing the policy space.

our two-types assumption regarding informedness, after learning occurs, all agents will be correctly informed.

3.3 Dynamics with Myopic Agents

We end this section by briefly noting the benchmark dynamics that would arise in a world with only myopic agents. Since myopic agents ignore the effect of learning on future outcomes, they express stage game preferences in each period. The first period policy will reflect the (original) median effective income earner's ideal policy, which we showed would belong to either the informed rich or the misinformed poor (i.e. $\tau_{med} = \max\{\tau_{LM}, \tau_{HI}\}$). If $\tau_{med} < \tau^{\dagger}$, then there will be no learning; the second period environment will be identical to the first, and the static outcome will repeat.

By contrast, if $\tau_{med} > \tau^{\dagger}$, all agents become informed, and then since the poor constitute a majority, the informed poor will be pivotal in the second period. Learning will cause political power to shift between the groups. Since the informed poor had the lowest effective income, and thus the highest ideal policy, taxes and public goods provision will be higher in the second period than the first. This is the slippery slope at work. A smaller equilibrium policy today begets a larger equilibrium policy tomorrow. There is endogenous policy momentum – the slippery slope. Because we assumed that agents were myopic, they did not attempt to manipulate policy to either prevent or ensure a slide down the slippery slope. We take up that concern in section 4.

4 Dynamics

In the previous section, we characterized the stage game preferences of agents, given their income and beliefs, and showed that these were single-peaked in the policy variable τ. Since the second period of our model is effectively a stage game, this characterization also reflects the agents' preferences in the second period. Additionally, since myopic agents are purely present-oriented, these also reflect the preferences of myopic agents in the first period.

The first period preferences of sophisticated agents may differ insofar as those agents understand that learning in the first period may affect second period outcomes. Absent learning, sophisticated agents understand that the equilibrium second period policy will reflect the ideal policy of the (original) median effective income earner, as characterized in the previous section, so that $\tau_2 = \tau_{med}$. By contrast, if there is learning, sophisticated agents understand that all agents will have the same second period beliefs and that the poor will be pivotal. Moreover, since all agents believe that *they* have correct beliefs, a sophisticated agent with belief A will assume that all other agents will arrive at that same belief.

A sophisticated agent's assessment of their lifetime utility given a first period policy τ is:

$$V(\tau; y, A) = \begin{cases} (1-\tau)y + A\ln(\tau\overline{y}) + \beta\left[(1-\tau_{med})y + A\ln(\tau_{med}\overline{y})\right] & \text{if } \tau < \tau^{\dagger} \\ (1-\tau)y + A\ln(\tau\overline{y}) + \beta\left[(1-\tau^{*}(y_L, A))y + A\ln(\tau^{*}(y_L, A)\overline{y})\right] & \text{if } \tau > \tau^{\dagger} \end{cases}$$

The first part of the agent's lifetime utility (compromising the first two terms) corresponds to first period utility. It is continuous and concave (and thus single-peaked) in the first period policy τ, and achieves a maximum at the stage game optimum $\tau^{*}(y, A)$. The second part corresponds to second period utility, and is affected by the first period policy τ only insofar as τ determines whether there is learning or not (i.e. whether τ is above or below $\tau^{\dagger}$). Thus, the second part is piece-wise constant in τ. If learning harms the agent, then there will be a discontinuous jump down in the agent's utility at $\tau = \tau^{\dagger}$. The opposite is true if learning benefits the agent. The size of this utility jump is given by $\Delta(y, A) = A[\ln(\tau(y_L, A)) - \ln(\tau_{med})] - (\tau(y_L, A) - \tau_{med})y$, which depends on both the agent's income and their beliefs, but not on the specific policy chosen in period 1.

Figure 1 illustrates the lifetime utility of a sophisticated informed rich agent as a function of the first period policy τ. Learning harms the informed rich, because it results in a second period policy that is farther from their ideal. Thus, learning reduces the agent's lifetime utility by $\beta\Delta$. The rightmost panel illustrates a situation where $\tau^{\dagger} > \tau_{HI}$; learning requires

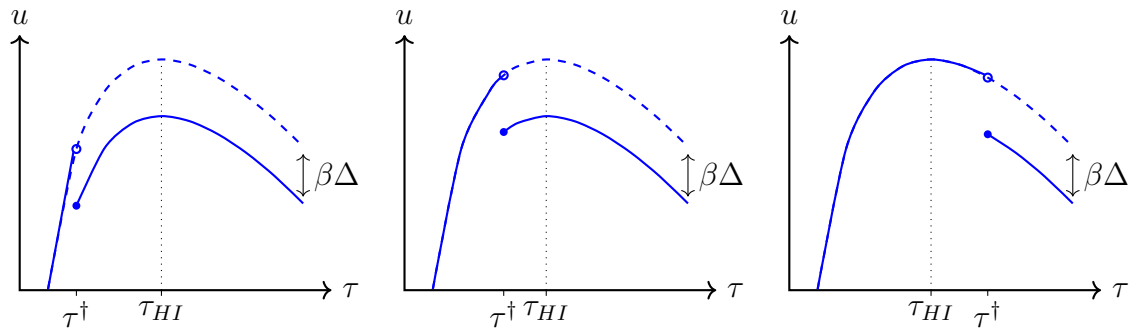

Figure 1: Lifetime utility for a sophisticated informed rich agent as a function of the first period policy τ. Each panel depicts utility for a different value of $\tau^\dagger$ — the threshold policy above which learning occurs; utility drops discontinuously at this threshold.

a first period policy above the agent's ideal stage game policy τ_{HI}. In this instance, the downshift in utility occurs in a region of the policy space where utility is already decreasing; single-peakedness is preserved. By contrast, in the other two panels, $\tau^\dagger < \tau_{HI}$, and so learning causes utility to decrease in a region where it is otherwise increasing, causing single-peakedness to be violated. Notice that, in the left-most panel, where a large distortion is required to prevent learning, the agent's most preferred policy remains unchanged; it is the stage game ideal. By contrast, her most preferred policy in the middle panel is $\tau^\dagger$. In this instance, the benefit from preventing learning exceeds the cost of distorting the policy away from the stage game optimum.

With these insights in mind, we are ready to begin characterizing the dynamic equilibrium of the game. Since the benefit of learning $\Delta(A, y)$ depends on which agent is pivotal absent learning (i.e. the informed rich or the misinformed poor), our analysis will be in two parts.

4.1 Divergence in Beliefs is Relatively Small

Suppose the divergence in beliefs between the informed and misinformed is relatively small (formally, that $\frac{A_I}{A_M} < \frac{y_H}{y_L}$). We previously showed that, in this scenario, $\tau_{HI} < \tau_{LM} < \tau_{LI}$, so that income disparities create a larger wedge in ideal policies than belief differences do.

If there is no learning, all agents expect the second period policy to be τ_{LM}. With learning, the sophisticated misinformed agents expect the same second period policy, anticipating that

all other agents learning that $A = A_M$. Thus, all misinformed agents will express stage game preferences in the first period, whether they are sophisticated or not. In particular, the most preferred first period policy of the misinformed poor will be their stage game ideal, τ_{LM}.

By contrast, sophisticated informed agents will potentially face dynamic incentives. They each understand that if there is learning, the second period policy will shift from τ_{LM} to τ_{LI} — which the informed rich perceive as worse, and the informed poor perceive as better. Hence, the informed rich may wish to strategically prevent learning that would otherwise happen under the myopic benchmark (i.e. if $\tau^\dagger < \tau_{LM}$), and the informed poor may wish to strategically induce learning when it would otherwise not (i.e. if $\tau^\dagger > \tau_{LM}$). However, neither group can build a majority coalition around such a strategic choice. For example, none of the poor agents (whether informed or not, or sophisticated or not) would join the informed rich in supporting a policy below τ_{LM}. Similarly, none of the other groups would join the informed poor in supporting a policy above τ_{LM}. This implies the following result:

Proposition 1. *Suppose the divergence in beliefs is small (i.e. $\frac{A_I}{A_M} < \frac{y_H}{y_L}$). The equilibrium first period policy is $\tau_1^* = \tau_{LM}$ (regardless of the value of $\tau^\dagger$).*

When the divergence in beliefs is small, behavior in the dynamic game coincides with the benchmark equilibrium with myopic players. The policy chosen in each period will simply be the most preferred policy of the median effective income earner in that period. In particular, the misinformed poor will implement their ideal stage game policy τ_{LM} in the first period. This will be true even if some measure of agents are sophisticated and have long run concerns.

If $\tau_{LM} < \tau^\dagger$, then there will not be any learning, and the first period outcome will repeat in the second period. If $\tau_{LM} > \tau^\dagger$, then there will be learning, and the informed poor will become pivotal in the second period. The second period policy will be $\tau_{LI} > \tau_{LM}$. There will be endogenous policy momentum.

When the divergence in beliefs is small, there may be policy momentum insofar as learning

causes political power to shift from the misinformed poor to the informed poor. However, awareness of this dynamic by sophisticated agents cannot sustain policy manipulation to prevent policy from going down the slippery slope.

4.2 Divergence in Beliefs is Relatively Large

Next, suppose that the divergence in beliefs between the informed and misinformed is relatively large (formally, that $\frac{A_I}{A_M} > \frac{y_H}{y_L}$). Then, $\tau_{LM} < \tau_{HI} < \tau_{LI}$; differences in beliefs create a larger wedge in ideal policies than income disparities do. If there is no learning, then the informed rich will be pivotal in the second period. By contrast, if there is learning, a poor agent will become pivotal. Unlike the previous case, in this setting, learning will be salient to all sophisticated agents, since it will be understood to shift political power from rich agents to poor agents. As before, learning hurts the informed rich since they perceive it as shifting political power away from them to the informed poor. For the same reason, the informed poor perceive learning as being favorable to them. Additionally, all misinformed agents perceive learning as favorable, since they perceive it shifting political power from the informed rich to the misinformed poor, whose ideal policy is closer to their own.

It turns out the dynamic incentives created by learning affect the strategies of the informed sophisticated agents quite differently from the misinformed sophisticated agents. Accordingly, we conduct the analysis in two parts. First, we suppose that all informed agents are sophisticated and all misinformed agents are myopic. Second, we suppose that all agents are sophisticated. A comparison of these cases will shed light on the role that sophistication plays in sustaining strategic behavior.[10]

4.2.1 Only Informed Agents are Sophisticated

Suppose that all informed agents are sophisticated and all misinformed agents are myopic. In this case, the strategic incentives are similar to those in Section 4.1. The misinformed will

10. A third alternative exists, where all informed agents are myopic and all misinformed agents are sophisticated. The equilibrium analysis in the second and third cases are quite similar. Accordingly, since this third case is seemingly the least empirically plausible, we leave the analysis to Appendix A.1.

express their stage game preferences. The informed rich may seek to strategically prevent learning that would otherwise happen (if $\tau^{\dagger} < \tau_{HI}$), and the informed poor may seek to strategically induce learning that would otherwise not (if $\tau^{\dagger} > \tau_{HI}$).

Similar to the previous section, the informed poor will not be able to build a majority coalition that successfully distorts policy; no other group desires policies above τ_{HI}. However, the informed rich may be able to successfully distort policy in a way that prevents learning. The reason is that, since $\tau_{LM} < \tau_{HI}$, the misinformed poor (who in conjunction with the informed rich constitute a majority) will support moves to push policy below τ_{HI}.

The willingness of the informed rich to distort policy depends on a comparison of the first period loss from the distortion against the second period gain from preventing learning. Naturally, the larger the required distortion, the less valuable it is to strategically prevent learning. Let $\underline{\tau}_{HI} < \tau_{HI}$ be the lowest first period policy (i.e. the most distorted policy) that prevents learning, that is acceptable to the sophisticated informed rich. We can easily verify that $\underline{\tau}_{HI}$ is the solution to:

$$1 - \left(\frac{\underline{\tau}_{HI}}{\tau_{HI}}\right) + \ln\left(\frac{\underline{\tau}_{HI}}{\tau_{HI}}\right) = \beta\left[1 - \frac{\tau_{LI}}{\tau_{HI}} + \ln\left(\frac{\tau_{LI}}{\tau_{HI}}\right)\right] \tag{2}$$

This condition precisely defines the point where the value to the informed rich agent for distorting in the first period and then achieving their myopic ideal in the second is equal to the value of playing the stage-game ideal in the first period and allowing learning to happen, with the value functions defined in Section 4.

Before stating the main result, we note a possible complication. The assumptions of the model do not guarantee that $\underline{\tau}_{HI} \geq \tau_{LM}$. If this condition is not satisfied (i.e. if the informed rich are willing to distort policy below the ideal policy of the misinformed poor) and if $\tau^{\dagger} \in (\underline{\tau}_{HI}, \tau_{LM})$, then the ordering of agents by their most preferred policy will differ between the static and dynamic games, affecting the identity of the pivotal voter. We begin by studying the case where this complication does not arise, and then subsequently address

the effect of the complication.

Proposition 2.A. *Suppose the divergence in beliefs is large (i.e.* $\frac{A_I}{A_M} > \frac{y_H}{y_L}$*), and that only informed agents are sophisticated. If* $\underline{\tau}_{HI} \geq \tau_{LM}$*, then the equilibrium first period policy is given by:*

$$\tau_1^*(\tau^\dagger) = \begin{cases} \tau_{HI} & \text{if } \tau^\dagger \leq \underline{\tau}_{HI} \\ \tau^\dagger & \text{if } \underline{\tau}_{HI} < \tau^\dagger < \tau_{HI} \\ \tau_{HI} & \text{if } \tau^\dagger \geq \tau_{HI} \end{cases}$$

The content of Proposition 2.A is summarized in the left panel of Figure 2. Implicitly, we assume that there is no learning when $\tau = \tau^\dagger$. If $\tau^\dagger \geq \tau_{HI}$, then there is no need for the informed rich to distort policy — implementing their stage game ideal policy is consistent with no learning, thus enabling them to retain power in the second period without any first period sacrifice. When $\tau^\dagger < \underline{\tau}_{HI}$, then the cost of distorting policy to prevent learning is so high that the informed rich implement their ideal policy today and accept that, by doing so, they will cede second period political power to the informed poor. Strategic manipulation occurs when $\tau^\dagger \in (\underline{\tau}_{HI}, \tau_{HI})$. In this case, the informed rich strategically under-provide the public good to prevent learning, enabling them to retain political power.

These insights can be restated in terms of the effectiveness of the policy feedback channel. When the feedback channel is weak (i.e. μ is high), then $\tau^\dagger$ will be large, and the likelihood of a slippery slope dynamic arising will be small. By contrast, when policy feedback is strong (i.e μ is small), then $\tau^\dagger$ will be small, and a slippery slope dynamic will be at play. When policy feedback is extremely strong, the distortion needed to prevent learning will be so high as to make strategic manipulation unattractive. By contrast, when there is moderate feedback, the informed poor strategically distort policy to prevent learning.

A comparison of Propositions 1 and 2.A — strategic behavior arises in the latter, but not the former — is instructive. Strategic manipulation will only succeed if there is a coalition

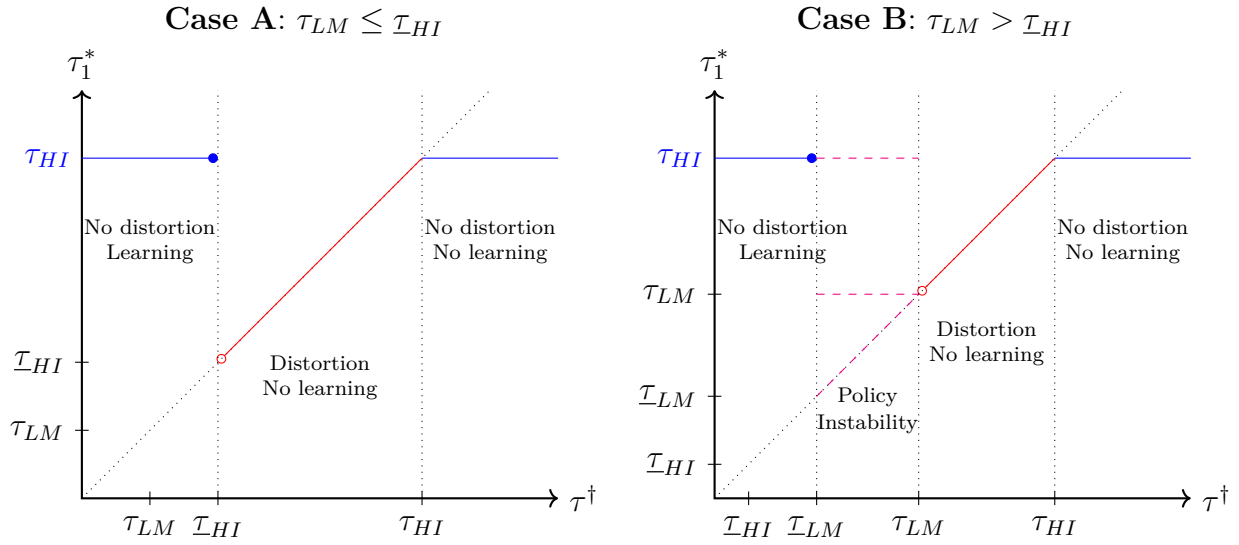

Figure 2: Equilibrium when the divergence in beliefs is large, the informed (only) are sophisticated. The left panel illustrates the case in Proposition 2.A where $\tau_{LM} \leq \underline{\tau}_{HI}$. The right panel illustrates the case in Proposition 2.B where $\tau_{LM} > \underline{\tau}_{HI}$. This includes a region of policy instability. In this region, the dashed lines represent the focal policies that generate the Condorcet cycle.

that will support it. When the ideal policy of the misinformed poor lies below that of the informed rich, a natural coalition exists that supports downwardly distorting policy. By contrast, when the ideal policy of the misinformed poor lies above that of the informed rich, then no such common incentive exists. Now, the ideal policy of the misinformed poor will be (relatively) low when the misinformed have strongly pessimistic beliefs. Thus, strategic manipulation is most likely when misinformation highly skews beliefs in the polity.

Let us now return to the complication that arises when $\underline{\tau}_{HI} < \tau^\dagger < \tau_{LM}$. In this scenario, the informed rich have an ideal policy ($\tau^\dagger$) below that of the misinformed poor (τ_{LM}), and this policy prevents learning; however, limiting attention to policies that induce learning, the informed rich have an ideal policy (τ_{HI}) above that of the misinformed poor. This preference order reversal causes policy to become unstable; there will be no majority winning policy. Instead, a Condorcet cycle will exist.[11]

11. In our baseline analysis, we associate the existence of Condorcet cycles with policy instability. In Appendix A.6 we explore more elaborate voting mechanisms that can fill these 'Condorcet Holes', and

Let $\underline{\tau}_{LM} < \tau_{HI}$ denote the policy below τ_{HI} that gives the misinformed poor the same utility as τ_{HI}.[12] (Though, in this section we focus on the case of $\beta = 0$, it will prove useful to define $\underline{\tau}_{LM}$ for any β.) We can easily verify that $\underline{\tau}_{LM}$ is the solution to:

$$\frac{\tau_{HI}}{\tau_{LM}}\left(1 - \frac{\underline{\tau}_{LM}}{\tau_{HI}}\right) + \ln\left(\frac{\underline{\tau}_{LM}}{\tau_{HI}}\right) = \beta\left[\frac{\tau_{HI}}{\tau_{LM}} - 1 + \ln\left(\frac{\tau_{LM}}{\tau_{HI}}\right)\right] \tag{3}$$

where, as above, this equation defines the policy where the informed poor receive the same value for a distorted policy followed by the informed rich agent's stage-game ideal that they do for the informed rich agent's stage-game ideal followed by the informed poor agent's ideal with learning. In fact, since in the current context, the misinformed poor are assumed myopic (i.e. $\beta = 0$), then we must have $\underline{\tau}_{LM} < \tau_{LM}$.

Proposition 2.B. *Suppose the divergence in beliefs is large (i.e.* $\frac{A_I}{A_M} > \frac{y_H}{y_L}$*), and that only informed agents are sophisticated. If* $\underline{\tau}_{HI} < \tau_{LM}$*, then the equilibrium first period policy is:*

$$\tau_1^*(\tau^\dagger) = \begin{cases} \tau_{HI} & \text{if } \tau^\dagger \leq \max\{\underline{\tau}_{LM}, \underline{\tau}_{HI}\} \\ \textit{No Majority Winner} & \text{if } \max\{\underline{\tau}_{LM}, \underline{\tau}_{HI}\} < \tau^\dagger < \tau_{LM} \\ \tau^\dagger & \text{if } \tau_{LM} < \tau^\dagger < \tau_{HI} \\ \tau_{HI} & \text{if } \tau^\dagger \geq \tau_{HI} \end{cases}$$

Proposition 2.B is summarized in the right panel of Figure 2, and is broadly similar to Proposition 2.A, except in that there is policy inconsistency over a region of the parameter space. To see why, note that, in this region, $\tau^\dagger$ cannot be a majority winner; the informed and misinformed poor (who together constitute a majority) would replace it with τ_{LM} (which implies learning). But τ_{LM} cannot be a majority winner, since the informed rich and informed poor (who together constitute a majority) would replace it with τ_{HI}. But a coalition of the informed rich and the misinformed poor would, in turn, replace this with $\tau^\dagger$ (thus preventing learning), provided that $\tau^\dagger$ is not too far below τ_{LM}. There is a Condorcet cycle.

predict a focal policy.

12. The notation intentionally highlights the analogy between $\underline{\tau}_{HI}$ and $\underline{\tau}_{LM}$. For $i \in \{HI, LM\}$, $\underline{\tau}_i$ denotes the policy below τ_{HI} that, absent learning, gives a type i agent the same utility as τ_{HI} does with learning.

We end this subsection by noting that, though we assumed that the informed poor were sophisticated, this assumption was not important to the result, and Propositions 2.A and 2.B would continue to hold if some or all of the informed poor were myopic. Similarly, it wasn't essential that *all* the informed rich are sophisticated. All that is required is that the measure of sophisticated rich agents is large enough that they, along with the misinformed (rich and poor) jointly constitute a majority.

4.2.2 All Agents are Sophisticated

Now, suppose that misinformed agents are sophisticated as well. The strategic incentives for the misinformed agents are quite different to the informed rich. Both types of misinformed agents will be willing to distort policy *upwards* to generate learning, anticipating that political power will shift from the informed rich (whom they consider to be optimistically misinformed) towards the misinformed poor (whom they consider to be correctly informed).

When $\tau^\dagger \in (\tau_{LM}, \tau_{HI})$, it may be the the informed rich will seek to prevent learning by locating policy just below $\tau^\dagger$, while the misinformed rich seek to induce learning by locating policy just above $\tau^\dagger$. Obviously, it cannot be that both groups successfully manipulate policy. Indeed, the informed rich will never prevail, since the informed and misinformed poor agents (who together constitute a majority) will always prefer a policy slightly above $\tau^\dagger$ that induces learning, to one slightly below it that does not. Since the informed rich cannot downwardly distort policy, we should never expect an equilibrium policy below τ_{HI}.

This does not guarantee that the equilibrium policy will be τ_{HI}. If $\tau^\dagger \leq \tau_{HI}$, the static equilibrium policy will generate learning, and so there is no additional incentive for the misinformed to strategically distort policy. By contrast, if $\tau^\dagger > \tau_{HI}$, then the misinformed may wish to upwardly distort policy, to induce learning that would otherwise not happen. Naturally, their willingness to do so depends on a trade-off between the first period cost of distorting against the second period gain from having political power shift in their favor.

Let $\overline{\tau}_{LM}$ denote the highest policy (i.e. the most distorted policy) acceptable to the misinformed poor that induces learning, assuming the myopic benchmark policy τ_{HI} does not. It is easily verified that $\overline{\tau}_{LM}$ is the solution to:

$$\frac{\tau_{HI}}{\tau_{LM}}\left(1-\frac{\overline{\tau}_{LM}}{\tau_{HI}}\right)+\ln\left(\frac{\overline{\tau}_{LM}}{\tau_{HI}}\right)=\beta\left[1-\frac{\tau_{HI}}{\tau_{LM}}+\ln\left(\frac{\tau_{HI}}{\tau_{LM}}\right)\right] \tag{4}$$

As before, let $\underline{\tau}_i$ denote the policy below τ_{HI} that, absent learning, gives a type i agent the same utility as τ_{HI} would with learning.[13] Recall that these expressions were defined previously by equations (2) and (3). Similar to Section 4.2.1, the equilibrium characterization will depend on the relative locations of $\underline{\tau}_{HI}$ and $\underline{\tau}_{LM}$.

Proposition 3. *Suppose the divergence in beliefs is large (i.e. $\frac{A_I}{A_M}>\frac{y_H}{y_L}$), and that all agents are sophisticated.*

1. *If $\underline{\tau}_{HI}\geq\underline{\tau}_{LM}$, then there exists $\tilde{\tau}$, with $\tau_{HI}\leq\tilde{\tau}<\overline{\tau}_{LM}$ such that the equilibrium first period policy is given by:*

$$\tau_1^*(\tau^\dagger)=\begin{cases}\tau_{HI} & \text{if } \tau^\dagger\leq\tau_{HI}\\ \tau^\dagger & \text{if } \tau_{HI}<\tau^\dagger<\tilde{\tau}\\ \textit{No Majority Winner} & \text{if } \tilde{\tau}<\tau^\dagger\leq\overline{\tau}_{LM}\\ \tau_{HI} & \text{if } \tau^\dagger>\overline{\tau}_{LM}\end{cases}$$

2. *If $\underline{\tau}_{HI}<\underline{\tau}_{LM}$, then the equilibrium first period policy is given by:*

$$\tau_1^*(\tau^\dagger)=\begin{cases}\tau_{HI} & \text{if } \tau^\dagger\leq\underline{\tau}_{HI}\\ \textit{No Majority Winner} & \text{if } \underline{\tau}_{HI}<\tau^\dagger<\overline{\tau}_{LM}\\ \tau_{HI} & \text{if } \tau^\dagger\geq\overline{\tau}_{LM}\end{cases}$$

The content of Proposition 3 is summarized in Figure 3. To make sense of Proposition 3, there are two sets of deviations to consider. If $\tau^\dagger<\tau_{HI}$, then there will be learning under

13. We can show that $\underline{\tau}_{LM}>\tau_{LM}$, and that the misinformed poor will prefer a policy $\tau^\dagger$ that prevents learning to policy τ_{HI} that induces learning whenever $\tau^\dagger\in[\tau_{LM},\underline{\tau}_{LM})$.

the static benchmark, and so the informed rich will seek to strategically under-provide the public good. If $\tau^\dagger > \tau_{HI}$, then learning will not occur under the static benchmark, and so the poor will seek to strategically over-provide the public good.

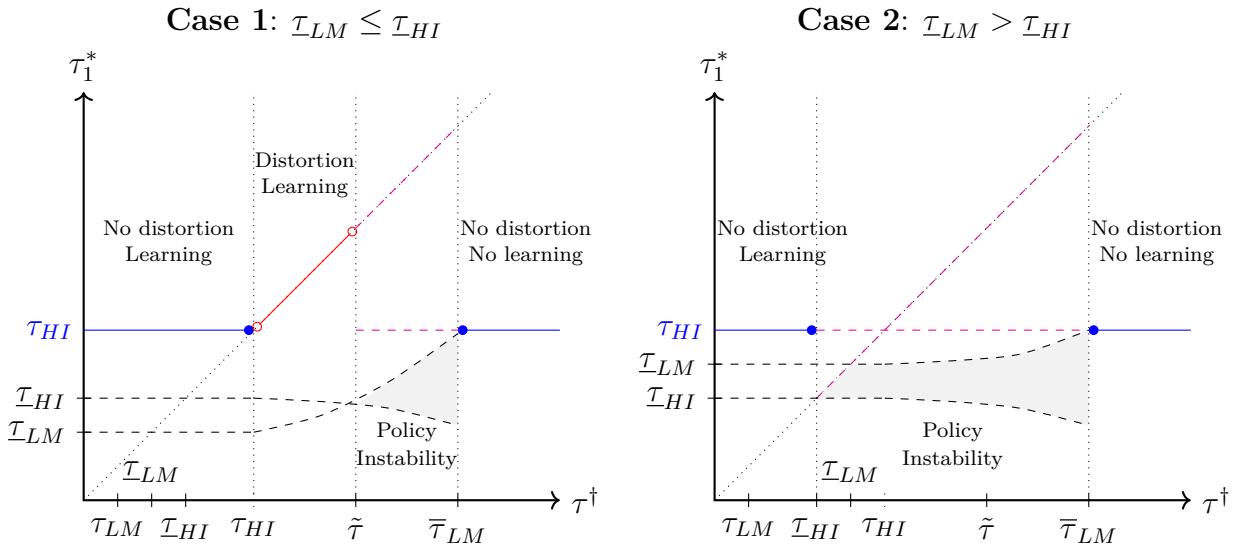

Figure 3: Equilibrium when the divergence in beliefs is large, and all agents are sophisticated. The left panel illustrates the case where $\underline{\tau}_{LM} \leq \underline{\tau}_{HI}$ and the right panel illustrates the opposite case where $\underline{\tau}_{LM} > \underline{\tau}_{HI}$. Both cases include regions of policy instability. The black dashed curves are constructed such that the informed rich and misinformed poor, respectively, are indifferent between the policies on their respective curves provided that these do not induce learning, and the policy policy $\max\{\tau^\dagger, \tau_{HI}\}$ (which does). The shaded region indicates policies that are strictly preferred by both groups to $\max\{\tau^\dagger, \tau_{HI}\}$. Whenever the shaded region is non-empty, a Condorcet cycle exists.

Let us take each motive in turn. Suppose $\tau^\dagger < \tau_{HI}$. By construction, the informed rich would prefer to replace τ_{HI} with a policy $\tau' < \tau^\dagger$ provided that $\tau' \geq \underline{\tau}_{HI}$. To be successful, they need the support of the misinformed poor. But since the misinformed poor benefit from learning, they will only support such a policy if it brings the first period policy much closer to their stage game ideal (i.e. if $\tau' < \underline{\tau}_{LM}$). If $\underline{\tau}_{LM} \leq \underline{\tau}_{HI}$ (as in part (1) of Proposition 3) then there is no policy τ' that satisfies both groups simultaneously, and so there is no possibility of successful downward distortion. If so, the equilibrium policy will be τ_{HI} whenever $\tau^\dagger < \tau_{HI}$.

By contrast, if $\underline{\tau}_{LM} > \underline{\tau}_{HI}$ (as in part (2) of Proposition 3), then any policy $\tau' \in (\underline{\tau}_{HI}, \underline{\tau}_{LM})$ that prevented learning would defeat τ_{HI} in a pair-wise majority contest. (These are the

policies in the subset of the shaded region in the right panel of Figure 3 satisfying $\tau^{\dagger} < \tau_{HI}$.) If so τ_{HI} cannot be a majority winner. But nor can any such policy τ' (assumed to be below $\tau^{\dagger}$) since a majority coalition of the informed and misinformed poor will replace it with a policy slightly above $\tau^{\dagger}$, and a coalition of the informed (rich and poor) would in turn replace *that* policy with τ_{HI}. A Condorcet cycle exists; there is policy inconsistency.

Next, suppose $\tau^{\dagger} > \tau_{HI}$. The misinformed poor would prefer to replace τ_{HI} with (a policy slightly above) $\tau^{\dagger}$ provided that $\tau^{\dagger} < \overline{\tau}_{LM}$, and in this endeavour, they will have the support of the informed poor. This policy is equilibrium consistent provided that it is immune to a counter-proposal by the informed rich that prevents learning and is closer to the stage-game ideal of the misinformed poor. Such a counter proposal will always exist when $\underline{\tau}_{LM} > \underline{\tau}_{HI}$ (for the reasons discussed above). Furthermore, even when $\underline{\tau}_{LM} \leq \underline{\tau}_{HI}$ a successful counter-proposal may exist if $\tau^{\dagger}$ is sufficiently large, so that the distortion required to induce learning is high. (The set of such counter-proposals is the subset of policies in the shaded regions in Figure 3, satisfying $\tau^{\dagger} > \tau_{HI}$.) The threshold $\tilde{\tau}$ is the largest distortion in policy that is immune to a counter-proposal. Whenever $\tau^{\dagger} > \tilde{\tau}$, a Condorcet cycle will arise and there will be policy instability.

A comparison of Proposition 3 against Propositions 2.A and 2.B reveals two key insights. First, when the misinformed are sophisticated, strategic manipulation produces the opposite dynamic to the typical slippery slope behavior. Rather than downwardly distort policy to prevent other agents from learning that a policy is more desirable, here, the agents upwardly distort policy to ensure that other agents learn that the policy is less desirable. This strategic behavior has a 'learning from mistakes' flavor to it — with the understanding that the mistake must be large enough to ensure that the perceived optimists learn their lesson.

Second, as more agents are made sophisticated, the possibility of coherent policy making breaks down, as strategic incentives cause disparate coalitions to pull policy in different directions. Furthermore, as we argued in previous sections, our results are robust to allowing

some agents from each group to be myopic. The insights in Proposition 3 did not rely on all agents being sophisticated *per se* — just that enough of them were.

Finally, recall that, to simplify the analysis, we assumed that the misinformed rich were never pivotal coalition partners. This made it unnecessary to consider policy deviations by a coalition of the informed poor and misinformed rich. Propositions 2.A, 2.B, and 3, will all continue to hold even if we relaxed this assumption, though the analysis would become more complicated.

4.3 Discussion

Given the preceding analysis, several insights become apparent. First, the slippery slope dynamic arises due to a specific interaction between the nature of misinformation and learning. It requires that: (i) misinformation causes the median agent to demand less of the public good than they would if perfectly informed; and that (ii) there is learning by acquaintance, so that beliefs evolve with the provision of the public good. The aggregated social preferences are time inconsistent, and reflect a shifting of political power between different groups of agents as learning occurs. Together, these features imply that, over time, the median agent's demand for the public good increases, creating an endogenous policy momentum whereby moderate policies today beget more extreme ones tomorrow.

We emphasize that the slippery slope dynamic is a political economy phenomenon — it is not enough that some voters be misinformed; what matters is how misinformation affects the aggregated social preference over outcomes. This insight will become particularly apparent in Appendix A.2, where misinformation causes agents to overvalue (rather than undervalue) the public good. There, we show that, although the demand by some agents may be higher, misinformation will not distort the preferences of the median agent. Hence, despite some agents being misinformed, and despite the possibility of learning by those agents, there will be no natural force causing policy to endogenously evolve.

Second, the possibility of a slippery slope dynamic arising may create incentives for particular groups of (sophisticated) voters to strategically manipulate policy. However, we showed that in a political economy setting, their ability to successfully do so is constrained by several factors, including the ordering of agents' stage-game ideal policies. Indeed, for there to be scope for strategic manipulation, the divergence in beliefs between the informed and misinformed needed to be relatively large. This ensured that the wedge in stage-game ideal policies arising from misinformation (between agents having the same income) was larger than the wedge arising from income differences (between agents having the same beliefs) —i.e. $\tau_{LM} < \tau_{HI} < \tau_{LI}$. This arrangement of ideal policies created the possibility that the group seeking to strategically manipulate policy could make common cause with other groups to build a majority coalition around the distorted policy. If this condition were not met, then the groups would seek to pull policy in opposite directions, preventing the emergence of a coherent majority coalition that could shift policy away from the stage-game baseline.

Third, we highlighted the crucial role that sophistication played in generating and sustaining policy distortion. We demonstrated that the canonical case, of policy under-provision to prevent a slide down the slippery slope, could only arise when informed agents were more likely to be sophisticated than their misinformed counterparts. This made possible a stable majority coalition between the informed rich and the misinformed poor to keep policy low. By contrast, when the misinformed agents were relatively more likely to be sophisticated, the opposite effect arose: the misinformed would upwardly distort policy to induce learning. Interestingly, the learning motive here had a 'learning from mistakes' flavor to it — the misinformed upwardly distorted policy by sufficiently much to teach their counterparts a lesson, by making it inescapably clear that the public good was not nearly so valuable.

Finally, we showed that sophistication amongst agents created the possibility of preference reversals, where the ordering of groups' ideal policies were different in the region of policy-space where learning occurred, from the region where it did not. We showed that these

preference reversals were associated with the existence of Condorcet cycles and incoherent policy making. Moreover, we showed that policy inconsistency became more likely as the number of sophisticated agents in the polity grew. This suggests that the role for actual strategic manipulation of policy motivated by a slippery slope dynamic is potentially quite limited. Though slippery slope arguments are common place as rhetorical devices, their translation to actual policy is necessarily more complicated.

5 Conclusion

Slippery slope arguments are ubiquitous in political discourse. In this paper, we explored a political economy mechanism that rationalized the slippery slope concern. We first showed that misinformation (that creates policy skepticism) combined with learning by acquaintance, can create a dynamic in which a small reform today begets larger reforms in the future.

We then examined whether awareness of the slippery slope dynamic would result in strategic manipulation of policy to prevent learning — i.e. whether slippery slope concerns would actually cause agents to scuttle otherwise welfare enhancing reforms. Though some agents may always wish to manipulate policy, in a political economy equilibrium with majority rule, we show that policy can only be successfully manipulated if two conditions are satisfied. First, the degree of misinformation must be large relative to the baseline level of political disagreement in the polity. Second, informed agents must be more likely to be sophisticated (and thus understand the slippery slope dynamic) than misinformed agents. If these two conditions are satisfied, then a coalition of the sophisticated informed rich, along with misinformed agents can conspire to strategically manipulate policy. While we focus on a public goods provision setting with rich and poor agents, our insights would apply to any policy setting where the core features of policy skepticism and learning by acquaintance arise.

We also explore other possibilities in the Appendix. When the misinformed are relatively more likely to be sophisticated than the informed, we get the opposite effect — policy skeptics

strategically over-provide the reform, to cause optimistic voters to learn that the reform is less worthwhile than they think. This behavior has a 'lesson-teaching' flavor, and generates the opposite dynamic — there is policy reversal rather than policy momentum. Additionally, we demonstrate that policy skepticism by the misinformed is crucial to the mechanism: when the misinformed are over-optimistic, policy does not evolve endogenously at all.

Though our model is simple and stylized, we believe that it captures important insights about the nature of decision making in a political economy setting. The robustness of our results to variant assumptions (see Appendix A) suggests that our insights will continue to hold in more complicated models. Amongst many issues worth investigating are the implications of relaxing various standard assumptions that our analysis takes granted, such as common knowledge assumptions. Certainly, there is scope for further theoretical development of the role of learning in a political economy setting, which we leave for future analysis.

We are grateful to the editor and three anonymous reviewers for their helpful feedback.

References

Acemoglu, Daron, and James A Robinson. 2001. "A theory of political transitions." *American Economic Review:* 938–963.

Alesina, Alberto, and Guido Tabellini. 1990. "A positive theory of fiscal deficits and government debt." *The Review of Economic Studies* 57 (3): 403.

Altman, Nancy J. 2005. *The Battle for Social Security.* Wiley.

Baccini, Leonardo, and Lucas Leemann. 2012. "Information and Voting-How Voters Update Beliefs After Natural Disasters." In *EPSA 2013 Annual General Conference Paper,* vol. 435.

Baker, Scott, and Claudio Mezzetti. 2012. "A theory of rational jurisprudence." *Journal of Political Economy* 120 (3): 513–551.

Battaglini, M., and S. Coate. 2008. "A Dynamic Theory of Public Spending, Taxation, and Debt." *The American Economic Review* 98 (1): 201–236.

Benabou, Roland. 2000. "Unequal societies: Income distribution and the social contract." *American Economic Review* 90 (1): 96–129.

Benabou, Roland, and Efe A Ok. 2001. "Social mobility and the demand for redistribution: the POUM hypothesis." *The Quarterly Journal of Economics* 116 (2): 447–487.

Bénabou, Roland, Davide Ticchi, and Andrea Vindigni. 2022. "Forbidden fruits: The political economy of science, religion, and growth." *The Review of Economic Studies* 89 (4): 1785–1832.

Besley, Timothy, and Stephen Coate. 1998. "Sources of Inefficiency in a Representative Democracy: A Dynamic Analysis." *The American Economic Review* 88 (1): 139–156.

Campbell, Andrea Louise. 2011. *How Policies Make Citizens: Senior Political Activism and the American Welfare State.* Princeton: Princeton University Press.

———. 2020. "The Affordable Care Act and mass policy feedbacks." *Journal of Health Politics, Policy and Law* 45 (4): 567–580.

Cook, Fay Lomax, Lawrence R Jacobs, and Dukhong Kim. 2010. "Trusting What You Know: Information, Knowledge, and Confidence in Social Security." *The Journal of Politics* 72 (2): 397–412.

Day, Nancy E, and Patricia Schoenrade. 1997. "Staying in the closet versus coming out: Relationships between communication about sexual orientation and work attitudes." *Personnel Psychology* 50 (1): 147–163.

———. 2000. "The relationship among reported disclosure of sexual orientation, anti-discrimination policies, top management support and work attitudes of gay and lesbian employees." *Personnel review* 29 (3): 346–363.

Dent, George W. 1999. "Defense of Traditional Marriage." *JL & Pol.* 15:581.

Dewan, Torun, and Rafael Hortala-Vallve. 2019. "Electoral Competition, Control and Learning." *British Journal of Political Science* 49 (3): 923–939.

Dewatripont, Mathias, and Gerard Roland. 1992. "Economic reform and dynamic political constraints." *The Review of Economic Studies* 59 (4): 703–730.

Fegley, Tate, and Ilia Murtazashvili. 2023. "From defunding to refunding police: institutions and the persistence of policing budgets." *Public Choice* 196:123–140.

Fernandez, Raquel, and Dani Rodrik. 1991. "Resistance to reform: Status quo bias in the presence of individual-specific uncertainty." *The American economic review:* 1146–1155.

Fox, Justin, and Georg Vanberg. 2014. "Narrow versus broad judicial decisions." *Journal of Theoretical Politics* 26 (3): 355–383.

Griffith, Kristin H, and Michelle R Hebl. 2002. "The disclosure dilemma for gay men and lesbians:" coming out" at work." *Journal of Applied Psychology* 87 (6): 1191.

Groseclose, Tim, and Nolan McCarty. 2001. "The politics of blame: Bargaining before an audience." *American Journal of Political Science:* 100–119.

Heidhues, Paul, Botond Koszegi, and Philipp Strack. 2018. "Unrealistic Expectations and Misguided Learning." *Econometrica* 86 (4): 1159–1214.

Herek, Gregory M, and John P Capitanio. 1996. ""Some of my best friends": Intergroup contact, concealable stigma, and heterosexuals' attitudes toward gay men and lesbians." *Personality and Social Psychology Bulletin* 22:412–424.

Herek, Gregory M, and Eric K Glunt. 1993. "Interpersonal contact and heterosexuals' attitudes toward gay men: Results from a national survey." *Journal of sex research* 30 (3): 239–244.

Hirsch, Alexander V. 2014. "Experimentation and Persuasion in Political Organizations." *American Political Science Review* 110 (1): 68–84.

Hirsch, Alexander V., and Jonathan P. Kastellec. 2022. "A theory of policy sabotage." *Journal of Theoretical Politics* 34 (2): 191–218.

Jacobs, Lawrence R, and Suzanne Mettler. 2018. "When and how new policy creates new politics: Examining the feedback effects of the Affordable Care Act on public opinion." *Perspectives on politics* 16 (2): 345–363.

Kang, Myunghoon. 2022. "Let presidents fail: Congressional deference to presidents as gambling on failure." *Research & Politics* 9 (2): 1–6.

Koch, Julianna, and Suzanne Mettler. 2012. "Who Perceives Government's Role in Their Lives?" *Working Paper.*

Kurtz, Stanley. 2003. "Beyond gay marriage." *The Weekly Standard* 8 (45): 26–33.

Lerman, Amy E, and Katherine T McCabe. 2017. "Personal experience and public opinion: a theory and test of conditional policy feedback." *The Journal of Politics* 79 (2): 624–641.

Nix, Kathryn. 2012. "Comparative Effectiveness Research Under Obamacare: A Slippery Slope to Health Care Rationing." *Heritage Foundation, Washington, DC.*

Parameswaran, Giri. 2018. "Endogenous cases and the evolution of the common law." *The RAND Journal of Economics* 49 (4): 791–818.

Persson, Torsten, and Lars E.O. Svensson. 1989. "Why a stubborn conservative would run a deficit: Policy with time-inconsistent preferences." *The Quarterly Journal of Economics* 104 (2): 325.

Piza, Eric L., and Nathan T. Connealy. 2022. "The effect of the Seattle Police-Free CHOP zoneon crime: A microsynthetic control evaluation." *Criminology and Public Policy* 21:35–58.

Raju, Manu. 2009. "Lieberman says no to Medicare buy-in." *Politico.* https://www.politico.com/story/2009/12/lieberman-says-no-to-medicare-buy-in-030553.

Roberts, Kevin. 1989. *The theory of union behaviour: labour hoarding and endogenous hysteresis.* Technical report. Suntory, Toyota International Centres for Economics, and Related Disciplines, LSE.

Roy, Avik. 2013. "Sen. Harry Reid: Obamacare 'Absolutely' A Step Toward A Single-Payer System." *Forbes.* https://www.forbes.com/sites/theapothecary/2013/08/10/sen-harry-reid-obamacare-absolutely-a-step-toward-a-single-payer-system/?sh=26ddf7e73af9.

Schattschneider, Elmer Eric. 1935. *Politics, pressures and the tariff.* Prentice-Hall, Inc.

Somin, Ilya. 2012. "A mandate for mandates: is the individual health insurance case a slippery slope." *Law & Contemp. Probs.* 75:75.

Strulovici, Bruno. 2010. "Learning while voting: Determinants of collective experimentation." *Econometrica* 78 (3): 933–971.

Volokh, Eugene. 2003. "The mechanisms of the slippery slope." *Harvard Law Review* 116 (4): 1026–1137.

World Public Opinion Poll. 2010. *American Public Opinion on Foreign Aid.* Technical report. World Public Opinion.Org, November. http://www.worldpublicopinion.org/pipa/pdf/nov10/ForeignAid_Nov10_quaire.pdf.

Biographies

Giri Parameswaran is an Associate Professor of Economics at Haverford College, Haverford, PA, 19041. Gabriel Sekeres is a Predoctoral Research Fellow at the Stanford Institute for Economic Policy Research, Stanford CA, 94305. Haya Goldblatt is a graduate of Haverford College, and currently unaffiliated.