Cliometrica (2021) 15:267–318
https://doi.org/10.1007/s11698-020-00207-0

ORIGINAL PAPER

What limits the efficacy of coercion?

Øivind Schøyen[1,2]

Received: 3 June 2019 / Accepted: 8 April 2020 / Published online: 10 May 2020

Abstract

We model a game between an authority, seeking to implement its state identity, and a parental generation, seeking to socialize a younger generation into their own identity. The authority first selects a coercion level against the non-state identity. The parental generation then chooses whether to insurrect in response to the coercion level and, if not, decides how much to invest in socializing their children into the non-state identity. In this overlapping generations model, we formalize and explore the consequences of an intrinsic negative reaction to coercion: coercion resentment. We show how coercion resentment can create an interval where coercion has negative efficacy in imposing the state identity. This causes the rational legitimacy maximizing authority to restrain its use of coercion. We then show how this inefficacy of coercion can make certain levels of coercion unimplementable without causing the non-state identity to insurrect. This causes the long-run equilibrium size of the non-state identity group to be dependent on their initial size and, thus, path dependence. We consider the validity of the model by reviewing two historical episodes: Stalin's secularization project (1922–1953) and the Counter-Reformation in early modern France and the Holy Roman Empire (1517–1685).

Keywords Coercion resentment · Political legitimacy · Identity · Insurrection · Path dependence

JEL Classification D02 · D10 · D82 · N30 · N40 · P16 · Z1

✉ Øivind Schøyen
oivind.schoyen@gmail.no

[1] International Research Fellow of Japan Society for the Promotion of Science, Hitotsubashi University, Tokyo, Japan

[2] FAIR Centre of Excellence, NHH Norwegian School of Economics, Bergen, Norway

1 Introduction

States generally seek to establish legitimacy—that is, to rule in accordance with the values of the ruled. Values and identities are inseparably linked, and many states build on particular state identities. The USSR built its legitimacy on the Communist identity and The Holy Roman Empire (HRE) on Catholic identity. One way of attaining legitimacy is by using extrinsic incentives—hereby referred to as coercion—to deter the spread of non-state identities, such as religious identities in the USSR and Protestants in the HRE.[1] The authorities of the USSR and the HRE were free to impose whatever level of coercion they wished. Yet, occasionally, if not always, they restrained their use of coercion, and, their non-state identities persisted. How can we understand the use and effect of coercion against non-state identities?

If people choose to internalize the identity that gives them the highest extrinsic utility, group identities will be a function of the institutional environment.[2] This would imply that everyone internalises the most opportune identity. In most cases, this would be the identity aligned with the ruling regime. The state's ability to instil its state identity by use of coercion would then be unlimited. If this was the case, the dynamics of religious, national, or ideological identities could be ignored in political economy analysis.

However, this understanding would poorly account for the following two stylized facts regarding state coercion against non-state identities: *Restrained coercion use:* coercion levels in consolidated authoritarian states with strong state capacity vary across states and within states over time. If coercion had unlimited efficacy, one would expect there to be little variation in the coercion level, and, that the coercion level would be persistently high in stable authoritarian states with high state capacity. *Varying persistence of non-state identities:* non-state identities in hostile institutional environments can be remarkably persistent, as demonstrated by the historical evidence of Jews in medieval Spain, as presented in Greif and Tadelis (2010). In contrast, other historical examples, like David Laitin's study of the Russian diaspora in the former USSR (Laitin 1998), reveal prompt adoption of new beliefs, norm sets, and national identities.

In order to account for these two stylized facts, we focus on the micro-assumption of an intrinsic counteraction to the use of coercion, which we refer to as coercion resentment. The paper embeds the assumption of coercion resentment and an insurrection risk on the use of coercion into Bisin and Verdier's (2001) overlapping

[1] In this paper, we imply the following when referring to "state legitimacy": the enforcement of a state rule by an authority $\mathbb{S}$ is legitimate if it sufficiently adheres to a set of internalized values s, such that an individual having internalized s does not feel obliged to resist $\mathbb{S}$. Further, when referring to identity, we imply a vector of beliefs and values that are internalized and embedded in a person—for example political ideologies or religious identities. We discuss binary identities that are transferable to any member within a population, e.g. not dependent on visible genetically inherited traits.

[2] This paper distinguishes between intrinsic and extrinsic utility. The intrinsic utility from an activity is the inherent reward from the activity, e.g. self-expression or enjoyment of the activity. The extrinsic utility of an activity is utility gained through the activity that can serve other purposes, e.g. money or status. This definition is taken from Kreps (1997).

generations (OLG) model of identities. In this model, we analyse a game between an authority seeking to maximize the prevalence of the state identity by using coercion, and, non-state identity parents seeking to socialize their children according to their own identity. The authority sets the coercion level whereupon the parents choose how much to invest into socializing their children into the non-state identity, and, whether or not to insurrect.

The model analyses how a rational legitimacy maximizing authority sets the coercion level when coercion triggers two opposing effects: the extrinsic incentives to comply and the intrinsic incentives to resist. The balance of these effects determines the prevalence of the non-state identity and, thus, their ability to insurrect against the authority. The paper develops four novel theoretical results. Here, we present historical evidence for these results by considering historical episodes that fit the premise of our model: Stalin of the USSR and the monarchs of early-modern France and the HRE. These authorities had clear and stated agendas to use the state apparatus to deter socialization of their respective non-state identities. They also lacked any institutional constraints on their use of coercion. The results of the model discussed below:

Restraints We show that if an authority seeking to maximize legitimacy restrains its use of coercion, it must imply a nonlinear coercion resentment response by the non-state identity parents. In order to assess the validity of this result, we review historical evidence from Stalin's secularization policies towards Christians and Muslims in the USSR (1922–1953). This secularization project was conducted in a comparatively cautious manner in regions where cultural differences were larger, recognizing the potential counterproductiveness of secularization attempts, thereby supporting the basic premise of the model.

Coercion dependence We show how authorities can be dependent on applying coercion to sustain their rule. This occurs if removing coercion sufficiently increases the prevalence of the non-state group, and, thus, their insurrection capability, thereby causing the non-state identity group to choose to commit an insurrection. We apply the analytic tool of coercion dependence to analyse the spread of Protestantism following Luther (1517). The introduction of the new protestant identity posed a direct threat to the unity of the HRE, which built its legitimacy on the Catholic faith. Separatist elements used the Protestant faith actively to deter the influence of the ruling Hapsburg elite, thereby posing an existential threat to the HRE state.

Dynamically stable equilibrium We show what combinations of non-state identity prevalence and coercion levels are stable in the long run. By long run, we imply a time interval such that the insurrection capability of the non-state group determined by their prevalence. The prevalence of the non-state group and, thus, the insurrection capability constraining the authority, is determined by the coercion level. Thus, setting the long run coercion level becomes a dynamic problem. We show that there are three unique classes of coercion levels that may constitute a dynamically stable equilibrium (DSE). We also show that every legitimacy maximizing authority will impose, or be in a process of imposing, a uniquely determined DSE. The French

crown experience of The Reformation can be understood as a move from a short-term equilibrium to a long-term DSE. The crown went from a non-confrontational policy of appeasement of the Protestants with the Edict of Nantes in 1589 to the gradual increase in pressure to adopt the Catholic faith, thereby making public worship for Protestants illegal, and finally imposing the death penalty for practizing Protestantism at the Revocation of the Edict of Nantes in 1685.

Path dependency We show how authorities with equal military technology and population responses to coercion can have different DSE. This occurs when different initial sizes of non-state identity groups give authorities different sets of coercion levels they might implement without causing an insurrection. Hence, the model displays path dependency: past shocks might affect what coercion levels are currently implementable, and thus the current DSE. We review comparative evidence from the Counter Reformation in early modern France and the HRE (1517–1685) to see if the mechanism of path dependency can explain the freezing pattern in the map of religious identities in Europe following the Peace of Westphalia in 1648.[3]

This paper is part of the literature on mathematical models explaining the spread of cultural traits. This literature began with Cavalli-Sforza and Feldman (1981), which pioneered the application of OLG models inspired by Darwinian biology to analyse the spread of cultural phenomenon. Boyd and Richerson (1988) extend this analysis by allowing for culture to be adopted from both parents and society in general. Bisin and Verdier (2000, 2001) focus on explaining why populations do not converge to a single identity equilibrium as children will be more likely to adopt a majority identity through socialization from society in general.[4] They explain equilibrium persistence of minority identities from the asymmetry in the cost of failed parental investment, thereby giving minority parents greater incentive to invest into socialization.[5]

Greif and Tadelis (2010) address identity persistence in the presence of a state authority seeking homogeneity. Greif and Tadelis (2010) account for the persistence of non-state identities in the presence of coercive states by focusing on the role of the

[3] This freezing pattern is thoroughly documented in political science literature (Tilly and Ardant 1975; Rae 2002; Nexon 2009; Johnson and Koyama 2019).

[4] As indicated by Gellner (2008), as societies become industrial, cultural homogeneity is the overall trend. Gellner (2008) explains this pattern with industrialization. Drawing mainly on empirical evidence from the European Early Modernity, he argues that the primary mode of economic cooperation dictates cultural identity. As industrialization made the nation-state the primary mode of cooperation, non-state identities vanished. However, as Gellner (2008) indicates, there is little evidence that we are converging towards a one-to-one relationship between states and identities. According to Gellner, the ratio of states to nations is at least one to ten. A large number of these non-state identities have stable size, for example, the many Catholics in European Protestant countries, the Indian Sikhs or the Muslims of modern-day Russia. Further, the erosion of non-state identities happens, not by even decay, but as a dynamic phenomenon with ebbs and flows. For example Allardt (1979) describes the resurgence of non-state ethnic identification in Europe in the late 1960s, and 1970s occurring against the general backdrop of cultural homogenization.

[5] This pattern of minorities investing more in socialization is backed up by empirical studies of minorities investing more in socialization. See Bisin and Verdier (2010) for a number of different empirical sources from large-scale surveys to data on religious school enrolment relative the prevalence of religion.

ability to conceal one's identity in public life to avoid incurring the cost of coercion. In their model, the efficacy of coercion is limited by that at a certain level of coercion, parents choose to conceal their non-state identity rather than incurring the cost of coercion. Further, they discuss the role of schooling and how the distance between values and policies can weaken the ability of institutions to propagate its values.

This paper extends the model developed by Bisin and Verdier (2000, 2001) and Greif and Tadelis (2010) by adding coercion resentment and analysing the insurrection constraint on coercion. The agents in the model of Greif and Tadelis (2010) are static in the sense that they do not intrinsically respond to coercion; in contrast, our model introduces an intrinsic negative reaction to the use of coercion. This enables our model to account both for a rational authority restraining its use of coercion under certain conditions while using high levels under others. Our model is also novel in that it analyses the non-state parent's insurrection decision. The insurrection decision is modelled by introducing an upper insurrection constraint on the authority's use of coercion dependent on the size of the non-state identity group. The insurrection constraint's interaction with the variation in the efficacy of coercion, arising from coercion resentment, enables the model to explain long-term dynamics in the relationship between the authority and the non-state group. We consider how authorities can be dependent on applying coercion to avoid insurrections and how path dependency arises.

This paper relates to four strands of the economic theory literature: social economics (Akerlof and Kranton 2000; Bisin et al. 2011; Carvalho and Koyama 2013; Carvalho 2013), group conflict (Sambanis and Shayo 2013; Acemoglu and Wolitzky 2014), state legitimacy (Johnson and Koyama 2013; Greif and Rubin 2014; Saleh and Tirole 2019), and path dependency in societal outcomes (Rubin 2011; Bisin and Verdier 2017; Besley and Persson 2019; Acemoglu and Robinson 2019). The paper also relates to the literature on nation-building (Alesina and Reich 2013) and optimal sizes of nations as a trade-off between homogeneity and economies of scale (Alesina and Spolaore 2005). In particular, this paper relates to Alesina and Reich (2013) and Alesina and Spolaore (2005) by developing micro-foundation of the limit of nation-building and developing formal prerequisites for when nations will contain two identities.

The remainder of this paper is organized in the following manner: Sections 1.1 and 1.2 discuss the main assumption of the model, legitimacy maximization, and coercion resentment. Section 1 extends the OLG model of Bisin and Verdier (2000, 2001), to include a legitimacy-maximizing authority, coercion resentment, and an endogenously determined insurrection constraint. The section then discusses the validity of the results in light of the historical evidence. Section 2 provides a concluding discussion. The appendices contain proofs, the baseline model of Bisin and Verdier (2000), and, further analysis of our model.

1.1 The search for legitimacy

A premise of the formal model is that authorities seek to maximize legitimacy. It is not claimed that this policy objective is always the prime directive for every authority. Clearly, authorities have additional, occasionally conflicting, agendas. Historical work generally confirms that legitimacy maximizing was an agenda for the

authorities we consider—the monarchs of the European early modernity (Johnson and Koyama 2019) and the leaders of the USSR (Froese 2008). We return to discuss this in more detail in the historical section. Here, we present a few arguments as to why maximization of legitimacy often can be an interesting baseline, *ceteris paribus*, assumption for understanding the coercive policies of authorities.[6]

To any authority, having a high level of legitimacy is desirable for a number of reasons: as Max Weber argues, it increases the probability of remaining in power. If the people of a state do not adhere to the states identity, the identity they adhere to can be used as a legitimization for overturning the state.

Another key motivation of states in building national, ideological or religious identities is to make populations respond in a manner that is emotionally related to the identity represented by the state. This is what makes religious and national identities powerful tools for authorities—the ability of internalized norms to invoke reactions that align the interest of the individual with the perceived interest of imagined national, political, or religious communities. Thus, legitimacy reduces costs and expands the possibility frontier of imposing policy (Greif 2008); in addition, it increases the willingness for altruistic behaviour, such as conscription (Levi 1997), or paying taxes (Levi 1999).

Finally, a population with homogeneous identities enables central policymaking (Tilly 1992); indeed, services like law and policing, hinge on and grow out of common sets of norms and values.

In the short term, the most obvious means to gain legitimacy is to take norms and values as given, and rule in accordance with the prevailing majority identity. To authorities in polities with heterogeneous identities, this implies making compromises between identities where they are incompatible, typically at the cost of reduced legitimacy (Johnson and Koyama 2013). However, states might enhance their legitimacy by increasing the proportion of the population with internalized norms similar to those on which the state builds its institutions. This can be done either by application of "sticks" (disincentives and coercion) or "carrots" (increasing the incentives of belonging to the authorities' identity). Other measures include increasing socialization and easing communication by creating common standards — that is through building of roads, language standardization, common school systems, and investing in common symbols.

We focus on the "stick" approach—coercion—and how it invokes an intrinsic negative reaction, thereby making it a potentially counterproductive measure. To the extent that "carrots", i.e. positive incentives, invoke a negative reaction amongst the members of the non-state identity, the analysis generalizes to authorities imposing positive incentives for adhering to their identity.

[6] Note that one could apply a behavioural argument for why state authorities seek to maximize the state identity—that is authorities could seek to maximize the prevalence due to an innate drive to make others think as oneself. Golman et al. (2016) argue that this is a general human tendency and could also be true for state authorities. If aligning others with one's beliefs is rational in becoming a state authority, we do not need to distinguish whether the motivation is intrinsically or instrumentally motivated, as long as the behaviour is adaptive. In settings where becoming an authority is competitive, and seeking to align others with one's identity is an adaptive strategy to become an authority, individuals who exhibit this behaviour will emerge successfully from the competition.

1.2 Coercion resentment

Coercion resentment is defined as the response of increasing behaviour that coercion intends to reduce. In the formal model we develop below, we narrow down this definition. Here, coercion resentment is the response of increasing socialization of the non-state identity as a reaction to state coercion. This section focuses on evidence of a micro-level coercion resentment.[7]

States applying coercion to deter socialization of non-state identities to attain assimilation can cause a backlash effect of increased mobilization in the non-state group (Tilly and Ardant 1975; Allardt 1979; Rokkan 1999; Fouka 2016; Fukuyama 2018). A study of socialization investment of German-Americans in the 1920's, Fouka (2016) shows how this backlash effect can increase socialization investments. Fouka (2016) studies the effect of German language restrictions in American schools following WWI integration of Germans into an American identity, comparing school districts in Ohio and Indiana that imposed language restrictions for exogenous reasons, to those that did not. She finds that such restrictions significantly *increase* the tendency of German-Americans to give their children German names to distinguish them from Americans. This among other plausible sources of higher socialization, caused the children in districts with language bans to be significantly more likely to marry other Germans. This indicates that choosing German names appears to be a successful strategy for these parents in increasing group cohesion. Thus, *increasing* the cost of having a German identity *increased* the investment into children holding a German identity. A parallel study, Fouka (2017), finds that Germans responded to discrimination from general society — measured by voting and violence motivated by nationality—by *decreasing* their investment into socialization. This happens as a response to *increasing* the cost of having a German-American identity.

Thus, German-Americans responded to targeted intentional measures by increasing socialization, while an untargeted general increase in cost lead them to reduce socialization. This could be explained by the coercion resentment response. Coercion resentment can be considered as irrational to the individual, in the sense that it causes the individual to increase a type of behaviour that becomes more costly for both the "sender" of the identity (parent) and "receiver" (child). The following section considers what motivates coercion resentment by examining evidence from behavioural sciences.

The response of coercion resentment could be an emergent phenomenon arising from the interaction between other fundamental human drives. Here, we review five types of motivation documented in the behavioural sciences that can interact to create coercion resentment:

Rejection of dominance hierarchies One way to understand coercion resentment is as a specific behaviour arising from a more general human tendency of the rejecting of hierarchy. State coercion attempts to silence the non-state identity group. For

[7] The response generally fits with much of the macro-level historical evidence reviewed in the following sections as well as the general findings of political scientist (Tilly and Ardant 1975; Allardt 1979; Rokkan 1999; Fouka 2016; Fukuyama 2018).

the non-state group, state coercion can be understood as a means of subduing their members to a privileged group: the state identity group. This type of subduing can trigger an resentment among the non-state group.

Humans have anti-hierarchical feelings, prefer reciprocal treatment, and are sensitive to the use of coercion (Woodburn 1982; Gintis and Van Schaik 2013). For an individual, being averse to coercion produces behaviour that has both costs and benefits. The benefits of an innate negative reaction to coercion are that it results in behaviour that makes it less likely to be taken advantage of. The cost of being averse to coercion is the cost of conflict and forgoing beneficial relations that are perceived to be coercive. Then, the aversion of humans to coercion must reflect the cost and benefit ratio in the environment of the evolutionary history of human populations. In general, humans are comparatively egalitarian relative to other primates. Woodburn (1982) proposes that an explanation for this is the comparatively low cost of killing conspecifics among humans compared to the cost for other primates. During the long period during which humans developed in small hunter-gatherer societies, every member had access to lethal hunting weapons and would-be coercive leaders had little way of protecting them from ambush. Therefore, coercive human leaders were easy targets for lower-ranking humans. Comparatively, the cost of killing a coercive chimpanzee leader would be much higher in terms of risk and effort for low ranking-chimpanzees lacking in efficient weapons. Hence, human leaders would have had to be comparatively more cautious in using coercion in our long period as hunter-gatherers. This made a general innate aversion to coercion less costly for early humans, and, thus, a more adaptive trait.

Reciprocity Coercion resentment can arise as a group-level version of the general trait of reciprocity (Bowles and Gintis 2011): the tendency to retaliate against hostile actions and reward beneficial actions. The assumed mechanism is that individuals that have internalized the coerced identity and feel that the authority has harmed their group, wish to punish the group associated with the coercion through activities aimed at stopping the influence of the authorities.

Salience of fulfilling internalized norms Coercion can increase the salience of acting in accordance with internalized norms. For non-state identity individuals, one means to alleviate the dissonance is confronting individuals of the state identity and curb the spread of their identity. In other words, an individual who has internalized a set of values will receive intrinsic utility from actively deterring the influence of an authority pursuing an agenda of opposing his values by increasing rather than decreasing socialization of the coerced identity.

Signalling pro-sociality Once coercion is imposed on an identity, defying the coercion and acting according to the coerced identity become costly, and can, hence, be used as a credible social signal of being intrinsically motivated. Our need to fulfil our internalized norms most likely arises from a need to signal pro-sociality, both to ourselves and others. Humans appear comparatively more interested in signalling pro-sociality, possibly arising from early humans participating in tasks that require reliance on cooperation, and magnified by the sum of collective efforts, such as big-game hunting and warfare (Jaeggi et al. 2010; Turchin 2016).

Parochial altruism As coercion towards the non-state identity increases, the authority will be considered as a threat to the non-state identity group. This increased external threat may invoke an emotional reaction that triggers a need to protect the group. This increases investment in social identity activities for individuals who have internalized the non-state identity. Thus, the presence of a threat to the group increases in-group identity and strengthens hostility towards the out-group. This out-group threat effect is documented to increase the number of different group-related behaviours, including increased investment in socialization (Huddy et al. 2013).

2 The model

We now extend the OLG model of Bisin and Verdier (2000). This model is concisely presented in "Appendix 1". It is recommended that readers who are unfamiliar with this model read this appendix. The Bisin and Verdier (2000) model builds on the premise that the steady state equilibrium level of identities within a population can be modelled considering how much parents invest into socialization. As parental decisions are based on the desirability of their child having the same identity as theirs, as opposed to another identity, the unique internal equilibrium is given as a function of this utility differential: Δu. In this model, we include an authority that can issue a penalty, referred to as coercion, for adhering to non-state identity. Furthermore, we make assumptions of how the agents respond to this coercion and analyze the use of coercion under exogenous and endogenous constraints to which different levels of coercion can be imposed. In order to focus on the implications of coercion resentment and its ability to explain the macro-phenomenon of persistence of non-state identities, we make several abstractions. We briefly elaborate on the consequences of three abstractions made in the model.

First, note that social network effects in the contagion of identities are omitted and we assume complete mixing of the population. In other words, we ignore the "neighbourhood" effect that the clustering of identities will skew socialization towards persistence of non-state identities. Formally, this implies that the probability of what identity a child adopts if parental identity fails is given by the prevalence of the identities in the total population, rather than in a social network structure; thus, it is equal for children of both identities. Including network effects would decrease the probability of non-state children adopting the state identity and would in itself be a short-term source of persistence. However, network effects fail to explain long-term persistence in a fully connected social network.[8]

Second, note that we assume that there is no cost of coercion. We follow Greif and Tadelis (2010) in assuming that the authority can impose coercion at zero cost, this is equivalent to assuming that cost equals benefits in terms of imposing the coercion.

[8] This general effect of contact as giving rise to persistence has been explored by Kuran and Sandholm (2008) and, more recently, Bisin et al. (2016).

There are clear costs to coercion—for example the enforcement cost of the coercion and creating enemies of potentially valuable allies. There are also potential benefits to coercion — wealth and influence: Influence is manifested in the form of taking resources from a non-state identity that can be turned into strategic gain by giving these resources to one's allies. Wealth is created by keeping the resources of the non-state identity as rent for the authority and its allies.[9]

Third, the model implicitly assumes atomized agents and abstracts from the dynamics of legitimacy caused by communities, organizational structure, framing or strategic interaction among elites. These implicit simplifications are justified as long as community leaders are equally good at maximizing their own influence by playing on salient identity cleavages. If both identities have leaders that frame situations equally well in terms of creating saliency, then the underlying potential for a cleavage will be the relevant mechanism at play. In other words, if one considers the "facts on the ground"—that is the actually given potential for action, the cards in the hands of the community leaders, then—if, on average, they play their cards equally well—the mechanisms in the model will be the driving factors.[10]

2.1 A model of a legitimacy maximizing authority

There is an authority controlling a state, defined as a monopoly on the employment of coercion, π, within the territory where the population is situated. This authority builds its legitimacy on s, the state identity, and wishes to maximize its prevalence by imposing coercion for adhering to ns, the non-state identity. The population consists of a continuum of agents who live in two periods, as a child at time t and as a parent at time $t+1$. There are two identities, $i \in \{ns, s\}$. Identities are mutually exclusive—a proportion q_t of the parent population holds identity ns at time t, while $1-q_t$ holds identity s. The timing is as follows: the authority sets the coercion level, π, parents then choose socialization investment after observing the coercion level, π and q_t. New generations of parents and children arrive until a steady state equilibrium (SSE) is reached. The SSE prevalence of the non-state identity is defined as $q^*(\pi)$. Thus, the utility maximization problem of the authority, $U^{\mathbb{S}}$, is as follows.

$$\max_{\pi} U^{\mathbb{S}} = \min_{\pi} q^*(\pi) \tag{1}$$

In order to maximize the prevalence of s identity, the authority sets the level of coercion π for adhering to identity ns. The coercion level can be interpreted as ranging

[9] For a study of how changes in economic factors can change equilibrium outcomes under costly coercion, see Dippel et al. (2016) and Carvalho and Dippel (2016). Although it could theoretically be possible that cost of imposing coercion alone limits the use of coercion, none of the historical sources applied in the historical cases examined focus on the cost of coercion alone as a limiting factor. All historical sources discuss efficacy in interaction with control capabilities, thereby supporting the basic premises of the varying efficacy of coercion and risk of insurrection as the relevant factors.

[10] For recent and forthcoming decision theoretic models of the role of the community leader, see Verdier and Zenou (2018) or Greif and Schøyen (2020) on moral authorities.

from low, such as social sanctions or issuance of fines for having identity *ns*, to high, such as criminal penalties; the maximum feasible level, $\pi_{\max}$, is referred to as a gun-point threat.

Identities are transmitted from one generation to the next. Transfer occurs either through parental socialization, from parent to child, or through oblique transmission: the influence of the general population. We denote the probability that parental socialization is successful τ^m, and the cost of socialization $H(\tau^m)$. If parent socialization fails, the child adopt an identity through a random individual of the population; hence, the probability of adopting *ns* identity is equal to its prevalence in the population. Parents choose τ^m to maximize expected utility by balancing the cost of parental socialization, denoted by the function $H(\tau^m)$, and the benefit of a higher probability of successful parental socialization. The smaller the prevalence of the *ns* identity is, the greater the loss by failed parental socializing and the greater investment in socialization will be. Let the utility of an i identity parent having an j identity child be denoted u^i_j. Parents balance the cost of investing into socialization by the benefit of having their children internalize their identity. Including the level of coercion, π, and resentment towards s identity caused by coercion, $C(\pi)$, into the parental utility function from Bisin and Verdier (2000) established in Appendix, we attain the following utility function U^i for parents.

$$U^{ns} = [\tau^{ns} + (1-\tau^{ns})q_t](\overline{u} - \pi) + (1-\tau^{ns})(1-q_t)(\underline{u} - C(\pi)) - H(\tau^{ns}) \quad (2)$$

$$U^{s} = [\tau^{s} + (1-\tau^{s})(1-q_t)]\overline{u} + (1-\tau^{s})q_t(\underline{u} - \pi) - H(\tau^{s}) \quad (3)$$

Following Bisin and Verdier (2000) we assume *s* or *ns* parents have a symmetrical utility loss if their child adopts an identity opposite to their own. We also follow them in assuming that the cost function of socialization follows the Inada assumptions—that is, as the probability of successful socialization approaches one, the cost of socialization approaches infinite. The first assumption this model adds to the Bisin and Verdier (2000) model follows from Greif and Tadelis (2010)—that the utility of having an *ns* identity child is lower when there is coercion.

Assumption 1 Parental empathy for coercion: The utility of having an *ns* identity child is $(u^i_{ns} - \pi)$.

Second, following the discussion in section 2.3, we assume coercion resentment, imposing coercion invokes a negative intrinsic reaction among the *ns* non-state identity parents—that is, they will have lower utility in having a *s* identity child.

Assumption 2 Coercion resentment: The utility to an *ns* parent of having a *s* identity child is $(u^{ns}_s - C(\pi))$.

We now derive the optimal levels of τ^m from (2) and (3), which are given by the FOCs.

$$(1 - q_t)(\Delta u - (\pi - C(\pi))) = H'(\tau^{ns}) \tag{4}$$

$$q_t(\Delta u + \pi) = H'(\tau^{s}) \tag{5}$$

A necessary condition for a stable interior SSE level is equal levels of investment, τ^i, in parental socialization of *ns* and *s* identity. If parents invest equally in socialization, they have equal marginal costs: $H'(\tau^{ns}) = H'(\tau^{s})$; thus, we can use (4) and (5) and the Inada assumptions on the cost of socialization, define some initial coercion level π_0 that yields an internal SSE, $\pi_0 : q^*(\pi_0) \in (0, 1)$, to establish the following lemma.

Lemma 1 Internal, unique, and reachable SSE's: *For all pairs of* $\{\pi, \Delta u\}$ *, two exterior SSEs exist. For some, but not all, pairs of* $\{\pi, \Delta u\}$ *a unique stable interior SSE exists, given by* $q^*(\pi) = \frac{\Delta u - \pi + C(\pi)}{2\Delta u + C(\pi)}$. *Imposing a coercion level* π' *corresponding to an interior SSE* $q^*(\pi') \in (0, 1)$ *from an initial interior SSE* $q^*(\pi_0)$, *will make q converge to* $q^*(\pi')$.

In order to illustrate the dynamics of the model, let us assume that at time t the initial coercion level is $\pi_0 = \underline{\pi}$ and the population is in an interior SSE with $q^*(\underline{\pi})$. Assume that the authority changes the value of π at $t + 1$ to $\overline{\pi}$, where $\overline{\pi} > \underline{\pi}$, and that the net effect of coercion for *ns* identity parents, $(\pi - C(\pi))$, is sufficiently increasing in the interval $[\underline{\pi}, \overline{\pi}]$, such that $q^{*'}(\pi) < 0$.

At $t + 1$, q remains unchanged but investment in socialization changes. The *ns* parents will now invest less in socialization as they have a net lower utility in having *ns* identity children. The *s* parents will invest more in socialization as the outcome of unsuccessful parental socialization—having an *ns* identity child—is less desirable to them. Socialization efforts now differ and q drops to $q_{t+2} < q^*(\overline{\pi})$ for the first generation presiding over the change in π. At time $t + 2$, parents will make the socialization investment decision with q_{t+2}, which is strictly smaller than q_{t+1}. Hence, *ns* identity parents will face a higher probability of their offspring having *s* identity through oblique socialization and will consequently increase their parental socialization. The level of the minority identity q_t will converge towards $q^*(\overline{\pi})$ until the SSE condition of $\tau^{ns} = \tau^{s}$, i.e. equal investment in parental socialization, is restored at the SSE with $q^*(\overline{\pi})$.

Now, assume that some other t', $\pi'' > \overline{\pi}$ is imposed such that *ns* parents prefer their children to have the *s* identity. This implies that *ns* parents stop socializing their children into the *ns* identity. Children with failed parental socialization will still adopt the *ns* identity with probability equal to $q_{t'+N}$, but the population will converge towards $q^*(\pi'') = 0$, with uniform convergence—that is $q_{t'+N} < q_{t'+N+1} \forall t$.

We now discuss the coercion resentment function, $C(\pi)$. The functional form of the coercion resentment function can be understood as a normalization of the effect of coercion resentment relative to the effect of coercion normalized to a unit scale — that is, assumed to be simply π. Thus, discussion of the net effect of coercion for *ns* identity parents can be centred around the coercion resentment function, $C(\pi)$. First, a few fairly unrestrictive functional form assumptions are made about $C(\pi)$,

most notably that coercion resentment is non-negative at no coercion, $C(0) \geq 0$ and that $C'(\pi) > 0$ more coercion implies more resentment. Formally, we assume:[11]

$$C(\pi) \text{ is a function of the } C^2 \text{ class, it is } C(0) \geq 0 \text{ and it has } C'(\pi) > 0, \text{ over the domain } [0, \pi_{\max}]. \tag{6}$$

Let us first assume the coercion resentment function is linear, $C(\pi) = K_0 + K_1\pi$ for some K_0, K_1. Taking the derivative of $q^*(\pi)$ with respect to π yields:

$$\frac{\partial q^*(\pi)}{\partial \pi} = \frac{(K_1 - 2)\Delta u}{(K_1\pi + 2\Delta u)^2}. \tag{7}$$

From (7), we see that using coercion as a means to change $q^*(\pi)$ would have no efficacy for $K_1 = 2$, negative efficacy for $K_1 > 2$, or efficacy at any level for $K_1 < 2$.

In the anecdotal evidence section from the Soviet Union secularization project (1922–1953), we argue that the use of coercion partially reduced religious identity; thus, we can rule out $K_1 < 2$. Similarly, we argue that we do observe restraints on the use of coercion during this period, defined as imposing a strictly lower coercion level than the highest possible. Hence, if Stalin was rational, and there was limited coercion use due to the inefficacy of applying higher levels, we can rule out a linear function with $K_1 > 2$.

For a convex or a concave coercion resentment function, the problem of setting the optimal coercion level will have a unique extremal point at the π' that solves:

$$\Delta u = \frac{2C(\pi') - \pi'}{2 + 3C'(\pi')}. \tag{8}$$

Assuming convex or concave functional forms yielding unique equilibria cannot explain restraints under certain conditions while also observing very high levels of coercion. In the historical section, we argue that we observe both restraints and high levels from a single authority: Stalin. Assuming that Stalin's coercive policies were rational, which appears to be a plausible assumption given his access to large-scale bureaucracies of analysts, this anecdotal evidence rules out a unique optimal level and a convex or concave coercion resentment response.

Having investigated the consequences of a linear, convex, or concave $C(\pi)$ function, we conclude that if coercion resentment is to account for both restraints on the use of coercion and variance in coercion levels, the $C(\pi)$ function must be nonlinear. Then, two assumptions of nonlinearity seem natural:

(1) *Initially accelerating coercion resentment* We assume that increasing the coercion level from no coercion to a level such that it becomes salient would produce a marginally increasing coercion resentment response. By salience implies the extent to which a difference captures the attention of a decision-maker. In this

[11] C^2 is the class of functions for which the first and second derivatives are continuously defined over the entire domain of the function.

case, a certain level of coercion will capture parental attention. This can be understood in the following manner: as the authority increases π, it goes from being perceived as representative of *s* identity individuals, which favours and endorses *s* identity, to being perceived as an enemy of *ns* identity individuals, with aggressive intentions of reducing the prevalence of *ns* identity individuals. This triggers both the parochial altruism motivation and reciprocity to confront the authority on behalf of the identity group itself and the values the identity adheres to. This implies that, initially, the coercion resentment will be a convex function.

(2) *Eventually decelerating coercion resentment* Once a high level of coercion is reached, the authority has clearly defined itself as hostile. An increase in pressure does not continue to increase marginal growth of coercion resentment, as there is habituation to coercive conditions among those in the non-state identity group. In other words, an resentment response is already triggered and additional coercion levels increase the response but with a lower second derivative. This implies that eventually, once coercion passes a certain threshold, the marginal increase in coercion resentment will sink.

(1) and (2) imply an initially convex eventually concave, "S-shaped", coercion resentment function. We focus our analysis on this functional form. After having established the results from this analysis, we indicate how the analysis can easily be generalized for nonlinear functional forms, where there are several salient thresholds causing multiple consecutive areas of effective and ineffective levels of coercion use.

We define a point $\hat{\pi}$ in the open interval, $\hat{\pi} \in (0, \pi_{\max})$ and assume that:

$$C''(\pi) = \begin{cases} > 0 & \text{for} \quad \pi \in [0, \hat{\pi}) \\ = 0 & \text{for} \quad \hat{\pi} \\ < 0 & \text{for} \quad \pi \in (\hat{\pi}, \pi_{\max}]. \end{cases} \tag{9}$$

Furthermore, we make the following assumption of the $C(\pi)$ function.

Assumption 3 Initially accelerating, eventually decelerating coercion resentment: The marginal utility loss because of coercion resentment approaches zero at the beginning and at the end of $[0, \pi_{\max}]$: $\lim_{\pi \to 0} C'(\pi) = 0, \lim_{\pi \to \pi_{\max}} C'(\pi) = 0$, and is strictly larger than two at least at one point, $\pi' \in (0, \pi_{\max})$: $C'(\pi') > 2$.

The last part of Assumption 3, the existence of a level of coercion that is strictly marginally ineffective, preceded and followed by marginally effective levels of coercion, is a crucial assumption on which the following results rest: variation in the marginal efficiency of coercion. With no variation in the marginal effectiveness of coercion—that is if all levels of coercion in $[0, \pi_{\max}]$ were marginally effective or were strictly marginally ineffective—the result would be trivial. The authority would either always apply the maximum level of coercion or never apply any coercion at all. This mirrors our previous discussion of linear functional forms.

We denote the coercion level giving this infimum as $\pi_{\underline{q}} \in (0, \hat{\pi})$, and refer to it as a non-confrontational level of coercion. Furthermore, we denote $\pi^e_{\underline{q}}$ to be the first coercion level larger than $\pi_{\underline{q}}$ that has $q^*(\pi)$ equal to the unconfrontational level:

$$\pi^e_{\underline{q}} \text{ is defined as a coercion level such that } \pi_{\underline{q}} < \pi^e_{\underline{q}} \text{ and } q^*(\pi^e_{\underline{q}}) \equiv q^*(\pi_{\underline{q}}). \tag{10}$$

There will always be a unique supremum value of $q^*(\pi)$ in the concave part of $C(\pi)$. We denote this level as $\pi_{\overline{q}} \in (\hat{\pi}, \pi_{\max}]$. Building on the assumption of an S-shaped coercion resentment function, (9) and Assumption 3, it follows that we can develop Lemma 2 by defining three classes of functional forms of $q^*(\pi)$.

Lemma 2 Coercion level-SSE functional form: *$q^*(\pi)$ is characterized by the following properties:*

(1) *a unique global or local maximum($\pi_{\overline{q}}$) and a unique global minimum($\pi_{\underline{q}}$)*

or

(2) *a unique global or local maximum($\pi_{\overline{q}}$), a local minimum($\pi_{\underline{q}}$) and a global, potentially unique, minimum ($\pi' \in [\pi^e_{\underline{q}}, \pi_{\max}]$)*

or

(3) *a global minimum ($\pi'' \in (0, \hat{\pi})$, where $q^*(\pi'') = 0$).*

In addition, there will always be a local or unique global maximum at $q^*(0) = \frac{1}{2}$. [12]

The properties of $q^*(\pi)$ are dependent on the size of utility loss for parents from having children with differing identities, Δu, and on the strength of the coercion resentment relative to the intrinsic effect of coercion. Class III) applies when Δu is sufficiently small and coercion resentment is sufficiently weak, such that a coercion level $\pi'' \leq \pi_{\underline{q}}$ gives $q^*(\pi'') = 0$. If $q^*(\pi_{\underline{q}}) > 0$, then either class II) or class I) applies, depending on the concavity of $C(\pi)$ in $(\hat{\pi}, \pi_{\max}]$. If $C(\pi)$ is sufficiently concave — such that $q^*(\pi_{\underline{q}}) > q^*(\pi_{\max})$—then class II) applies; if not, then $\pi_{\underline{q}}$ is a global minimum, and I) applies. Note that class I) is qualitatively similar to a convex $C(\pi)$: it has a unique nonzero minimum $q^*(\pi)$ value. Class III) is qualitatively similar to a linear coercion resentment function—that is $C(\pi) = K_0 + K_1\pi$—with $K_1 < 2$, while II) is qualitatively non-convex. Figure 1 illustrates the three possible classes of $q^*(\pi)$.

We now focus on how the authority will set the coercion level. Several factors external to the model can constrain the use of coercion by an authority: the authority might recognize constitutional legal rights or there might be institutionalized rights constraining what π the state apparatus can issue. In order to analyse optimal

[12] When $q^*(\pi)$ is characterized by III), it may also have a unique global or local maximum ($\pi_{\overline{q}}$).

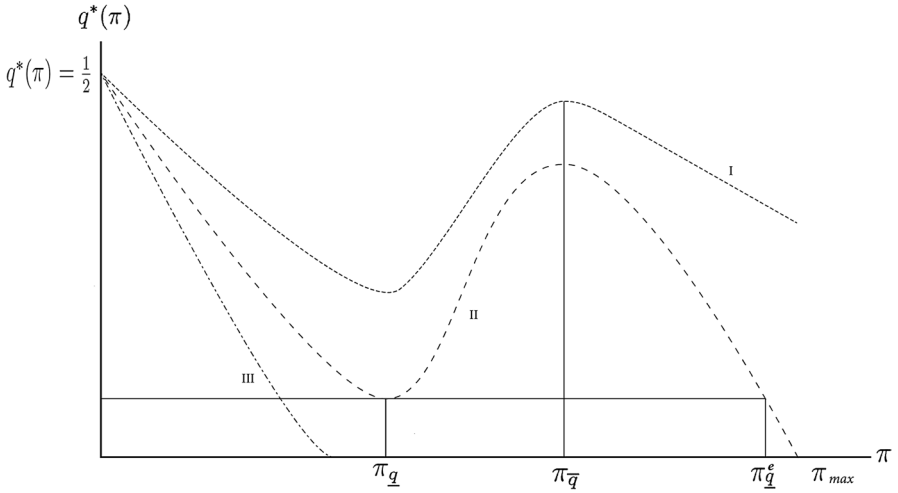

Fig. 1 Prevalence of the non-state identity, $q^*(\pi)$, as a function of the coercion level π. *Note:* Three examples of $q^*(\pi)$ from $\pi_0 = 0$, constructed using $C(\pi) = \tan^{-1}\pi$. The area between $\pi_{\underline{q}}$ and $\pi^e_{\underline{q}}$ constitutes the inefficient interval of coercion where coercion increases $q^*(\pi)$

use of coercion when the authority's ability to impose coercion is limited, π^{EC}, an exogenous constraint $\rho \in (0, \pi_{\max})$ is introduced. The maximization problem for an authority is the given by the following equation.

$$\max_{\pi} U^{\mathbb{S}} = \min_{\pi} \left[\frac{\Delta u - \pi + C(\pi)}{2\Delta u + C(\pi)} \right] \text{ s. t. } \pi \leq \rho \tag{11}$$

Trivially, an exogenous constraint $\rho \in (0, \pi_{\max})$ affects the optimal level of coercion π^{EC} if, and only if, it is strictly smaller than the optimal adjustment under no constraint, π^{NC}. Further, a rational authority will never impose a coercion level in the inefficient interval of coercion, $(\pi_{\underline{q}}, \pi^e_{\underline{q}})$. Noting this, we can develop the following lemma on the optimal level of coercion, π^{EC}, for an authority facing a constraint on the use of coercion.

Lemma 3 Coercion use under constraints: *If a constraint affects coercion use under an exogenous constraint,* $\rho \leq \pi^{NC}$ *and* $\rho \neq \pi^e_{\underline{q}}$*, the following holds.*

(i) $\pi^{EC} = \rho$ *if and only if* $\rho \notin (\pi_{\underline{q}}, \pi^e_{\underline{q}})$.
(ii) $\pi^{EC} = \pi_{\underline{q}} < \rho$ *if and only if* $\rho \in (\pi_{\underline{q}}, \pi^e_{\underline{q}})$.

Considering Lemma 2 on functional form, we see that the inefficient interval is only defined for functional form class II). We define an authority that imposes a coercion level strictly lower than its highest implementable level towards a non-state identity as exhibiting restraint. We established in the discussion that if both parts of Assumption 2—initially accelerating, eventually decelerating coercion resentment—is relevant for coercion efficacy, then $q^*(\pi)$ will be of class II). Thus, the efficacy of

coercion varies as a result of coercion resentment—going from having efficacy to having negative efficacy and back to having efficacy—only under class II). Building on these two definitions of variation and restraints, we can apply the optimization principle to establish the following theorem.

Theorem of Restraints on Coercion: A legitimacy-maximizing authority will restrain its use of coercion towards a non-state identity as a response to a constraint if and only if the efficacy of coercion varies and the restraint is in an inefficient interval of coercion.

We have established that, if an innate reaction to coercion can explain the macro-phenomenon of restraint, it must logically imply a nonlinear $C(\pi)$ causing the efficiency of coercion to vary. In order to see if it is plausible to consider restraints, we review evidence from the secularization project in the USSR on restraints before we return to the formal model.

Restraints on coercion and the Soviet secularization project 1922–1953

Stalin ($\mathbb{S}$) had a clear and stated agenda to reduce the prevalence of religious identities (ns) and used coercion (π) against the major religions of the USSR—Christianity and Islam—in order to increase the prevalence of its own secular identity, communism (s). This was both an expressed goal and part of the policy goals for most authorities of the USSR as well as a policy implemented during Stalin's reign (Lenin 1909; Kula 2005).[13] The approach towards religious communities in the USSR was initially very oppressive. The Great Terror of the 1930s witnessed widespread killings and forced gulag encampment of religious individuals who failed to denounce their religion. From 1937 onward, the Soviet authorities altered their policies towards religion. According to the Soviet authorities' own 1937 consensus, the coercive measures proved inefficient. The inefficacy of the measures combined with the need to apply religious and national sentiments at the beginning of the second world war (WWII), moved secularization measures from severe and strongly coercive, to unconfrontational and less malignant (Froese 2008).[14] This supports that, assuming Soviet authorities were accurate in their assessment that coercion was ineffective towards religious groups, coercion resentment is not of negligible effect—that is a linear coercion resentment function with $K_1 < 2$.

Data from Froese (2008) show how religious identity (q) in the USSR decreased as a consequence of deliberate Soviet policies to reduce its prevalence, while it increased again after the fall of the Soviet Union following the cessation of anti-religious policies (π) (see Figs. 2, 3). Overall, the attempt to secularize the Christian regions of Soviet society was successful in that it led to a drastic reduction in the

[13] Implicitly, we here assume that communism can be understood as a set of internalized values on par with religion. Indeed, the Soviet authorities themselves saw it this way Kula (2005).

[14] Illustrative of the approach of the authorities are the names of the atheist movement founded by the USSR authorities. Before 1922, the organization of atheists was named League of Militant Atheists, literally translated from Russian: League of the Militant Godless. It was disbanded at the onset of WWII. A subsequent atheist organization founded after WWII was named the Knowledge Society.

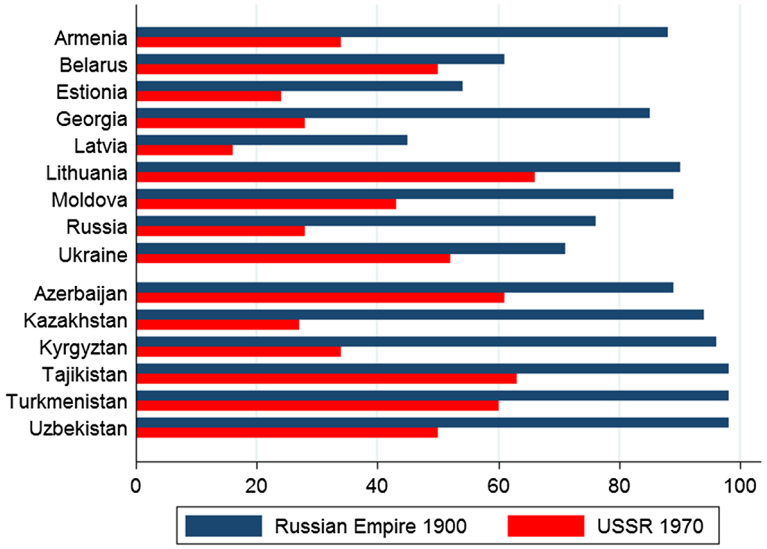

Fig. 2 Religious affiliation in the Russian Empire (1900) and the USSR (1970). *Note:* Religious affiliation with all religions before the Soviet Union, during the Russian Empire (1900), is represented by the bottom blue bars and in the Soviet Union (1970) is represented by the top red bars. All Christian countries have predominantly Orthodox Christianity as the majority religion, except Lithuania which is Roman Catholic and Latvia and Estonia which is Lutheran; all Muslim countries in the Soviet Union predominantly have Sunni Islam as their majority religion, except Azerbaijan, where Shia Islam is practiced (color figure online)

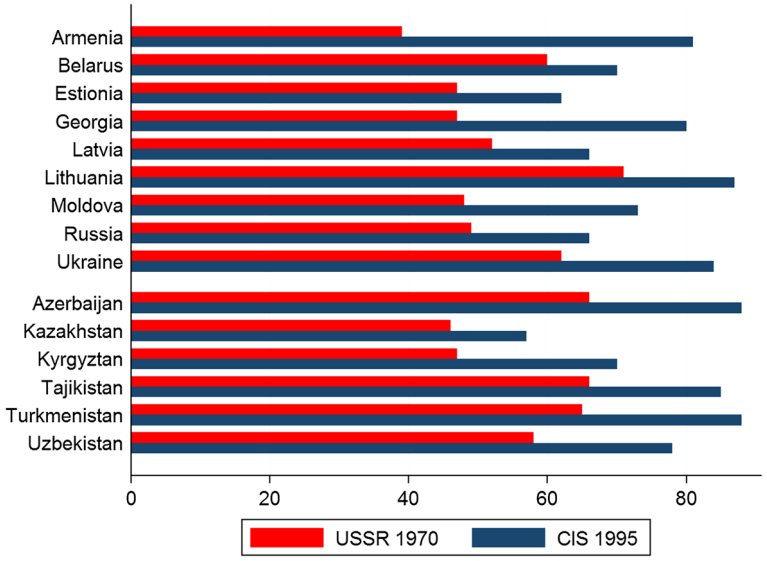

Fig. 3 Religious affiliation in the USSR (1970) and the Commonwealth of Independent States (1995). *Note:* Affiliation with the majority religion in the Soviet Union (1970) is represented by the top red bars and after the Soviet Union in the Commonwealth of Independent States (1995) is represented by the bottom blue bars. *Source:* Froese (2008) (color figure online)

prevalence of religious identity, but it did not lead to complete elimination of religious sentiment. Thus, if we accept that there is a causal link between the hostile institutional framework against religion in the USSR and the reduction in religious affiliation, we can reject a linear coercion resentment function with $K_1 > 2$.

From 1937 onward, the major Christian churches of the Soviet Union were able to continue their practice, although they faced censorship and demands from the authorities to serve the purposes of the Communist Party. The persistence of religion was stronger in areas where churches were aligned with other cleavages relative to the Russian amalgamate of identities associated with the Soviet rule. This was particularly true where the church was considered to be opposing the state. One example is membership of the autonomous Lithuanian Roman Catholic Church, which was considered as synonymous with resistance to the Soviet authorities (Clemens 2009).

The Baltic states had recently won their independence from Imperial Russia in 1918 before they were again annexed in 1940 by the USSR.[15] The desire for national independence interacted with the desire of religious freedom. This suggests that where the framing of religious persistence was aligned with other in-group and out-group dynamics, the Δu towards secular USSR identity was higher. A possible explanation is that, as the framework predicts, a higher Δu leads to a higher q_t, which gave members of these communities a higher utility in rejecting the authority in terms of social recognition and led to more visible resentment towards anti-religious policies.

In Central Asia, there were larger cultural differences between the local Muslim identity and the secular Communist identity, Δu. While most of the USSR had an ethnically Slavic, Christian population working in a mostly agricultural economy, Central Asia had ethnically Turkic, Muslim population with an economy based on trade and pastoralism . Adding further complexity was the recently nomadic Kyrgyz and Turkmen (Rywkin 1990; Wheeler 1967; Edgar 2004).

However, the potential gains for the local population from Soviet rule, modernization, and economic development, were also higher than that in the Baltic regions. Hence, in accordance with predictions from the classic Nash bargaining model (Nash 1953), Communist and local leaders on both sides had poor outside options and better incentives for cooperation, thereby contributing to a climate of communication between elites that was comparatively more benign than that between Moscow and Baltic elites (Kirby 1977).

The Soviet authorities framed communism as modernization, sweetening the deal of Soviet rule with promises of economic development to gain the partial support of Islamic intellectuals in Central Asia (Northrop 2001; Edgar 2004).[16] Policies

[15] This happened as a consequence of the secret protocols of the 1939 Molotov–Ribbentrop non-aggression pact between Nazi-Germany and the USSR.

[16] Froese (2008) describes how Soviet and Muslim authorities found common ground. Although the Communist agenda, in the long run, was to destroy Islam, which they saw as prejudice against reason, Lenin described "Muslim folk heroes as emblematic of the human struggle against oppression", while Muslim scholars noted that Islam could justify "even the rule of a usurper as means of assuring the public order and the unity of all Muslims". The tone between the Soviet and Muslim authorities can be read between the lines in a letter from the Central Religious Muslim Board in 1942 to Stalin, "...champion of the liberation of oppressed peoples and a man ever attentive to the need of the peoples...May Allah bring your work to a victorious end." (Marshall et al. 1971). Implicitly, the council signalled that they were sympathetic to Stalin's cause, but that he will not succeed without the assistance of Allah (Froese 2008).

such as the removal of Muslim courts were cautiously framed as modernization and done in co-operation with moderate Islamists. Muslim schools where not banned but encouraged to change their curriculum to teach science and employ Soviet trained teachers (Edgar 2004).

Initially, Stalin allied with Muslim modernization movements, most notably the Jadidism movement, which sought to "rationalize Islam, to purify it and bring it into line with the modern era" through "progress, development and growth". Although the secularization of the Central Asian USSR was deliberately non-confrontational, there were, however, clear reactions to the Soviet anti-religious policy. An illustrative example of this is the violent reactions to the 1920s Hujum policy of having Muslim women remove their veils (Northrop 2001; Edgar 2004).

Stalin would subsequently deceive his former Jadid allies and purge most of its leadership on suspicion of their ambitions for further national independence for the Central Asian republics (Bennigsen and Lemercier-Quelquejay 1967).[17] Stalin's fears of growing demands for autonomy were not unfounded. Bennigsen and Lemercier-Quelquejay (1967) describe how Muslim national identities that were barely present in 1917 emerged to gain increasing salience in at the time of publication in 1967. This happened, in part, as a result of the anti-religious policies and, in part, by the Soviets own policies to promote national identities.[18] Bennigsen and Lemercier-Quelquejay (1967) account for this effect mainly by what they describe as "resentment against cultural and administrative domination of the Russians". These sentiments could, potentially, be turned into momentum for an insurrection against the USSR. While the treatment of Muslims in Central Asia was relatively benign, the treatment of smaller groups of Muslims in the south-western region of Russia, such as the Crimean Tartar and the Chechens, was much more coercive and confrontational: forced deportations and subsequent expropriation of land were the primary instruments (Conquest 1970).

Looking across regions, Stalin persecuted either in a heavy-handed manner or maintained a non-confrontational approach (Froese 2008; Conquest and Case 1991). An explanation of the comparative differences in policies may be that the Soviet authorities were aware of an inefficient interval of coercion $(\underline{\pi_q}, \underline{\pi_q^e})$ and restraining their use of coercion as a response to an insurrection constraint in Central Asia, while pursuing levels beyond $\underline{\pi_q^e}$ in Europe, where they had no strategic constraints.

[17] However, the promises of development and growth were not reneged by the Soviet authorities. They trusted that their Muslim counterparts would not attempt to secede sufficiently to endow them with a better working economy. This growth happened alongside positive social changes in Central Asian USSR. For example, women were given a comparatively independent role and educational levels were increased, further integrating Central Asian USSR with Moscow (Edgar 2004). Together with the strengthening of Russian military capability, these changes made any threat of cessation less realistic (Conquest 1970).

[18] In his own work on ideology, Stalin was open to the theoretical possibility of "the right of secession" for national minorities (Stalin 1975). Stalin's policies also encouraged and facilitated formation of national identities in previously tribal areas in Central Asia. However, in practice, he considered any real transfer of power to the local level as highly undesirable, as it would give power to religious leaders which he saw as backward and counter-revolutionary (Conquest and Case 1991). This supports the hypothesis that Stalin's motivation for homogenization was instrumental rather than ideological. Creating national identities was a means to modernize the region and enable centralized control, thereby functioning to end the exertion of political power.

Assuming that Stalin perceived the response to coercion as stable across regions, this historical evidence supports the theorem of restraints on coercion and indicates that the combination of cultural differences and a negative response to coercion was sufficiently hostile as to be grouped under class II) in the USSR.

For now, we have taken the superior coercive capabilities of the USSR relative to the Muslim leaders as given. Beyond having availability of vastly better technology, in the widest use of the term, one crucial reason for the Communist ruling the state was that individuals who had internalized the Communist identity had higher prevalence within the USSR population. Indeed the Communist had just recently ceased power in a revolution following depletion of the legitimacy of Imperial Russia. The spread of Communist identity leading up to the Russian revolution obviously contributed to the Communist capability to take over of Imperial Russia. Considering Russian history, it appears reasonable to assume that a larger prevalence of state identity leads to a larger coercive capability and lower insurrection risk for a state authority. Conversely having a large part of the population adhering to a non-state identity would tend to limit a state's capacity for coercion. Further, it will tend to increase the risk of insurrection by the non-state group as a response to state coercion. We now extend our model by making the insurrection constraint dependent on the prevalence of the non-state identity to account for this.

2.2 Endogenous insurrection constraints and coercion dependence

We now analyse the model including the non-state parents' decision on whether to respond to a coercion level by instigating an insurrection decision. We consider the infinitely repeated game, where the authority first selects π', whereupon *ns* identity parents collectively choose whether to insurrect, and if not, parents of both identity groups select their levels of socialization investment τ^{ns}, τ^{s}. New generations of parents then continue to set the socialization investment for each generation. When the population is in the new SSE, $q^t = q^*(\pi')$, the authority can set a new coercion level, and the game is repeated.

We assume the parents' decision to commit an insurrection is dependent on their ability to do so. This capability is affected by a number of factors external to the model: the availability of military capital and strategic positions and partners. However, it appears natural to assume that, *ceteris paribus*, these external conditions, any given non-state identity group will have a smaller capability to insurrect given that they shrink in size. Conversely, it appears natural to assume that a non-state identity group will have greater capability to insurrect given that they grow in size. Thus, their capability to insurrect is assumed to be positively related to the SSE size of the non-state group, $q^*(\pi)$.[19] Since this size of the non-state identity group is given by the coercion level, as established in the previous section, the constraint on coercion is now an endogenous insurrection constraint on the use of coercion.

[19] For tractability, the insurrection constraint $\rho(\cdot)$ is assumed to be dependent on $q^*(\pi)$ rather than on q_t: $\rho(q^*(\pi))$. This assumption is not very restrictive and the properties of alternative approaches are discussed in "Appendices 3.2 and 3.3".

The insurrection constraint is defined as the highest coercion level for which the minority has the negative excepted utility of committing an insurrection.[20] The insurrection constraint function $\rho(q^*(\pi'))$ defines the maximal coercion level that can be implemented for an SSE $q^*(\pi')$ without the *ns* identity committing an insurrection. Note that there is no explicit link between the insurrection decision and coercion resentment. The private decision of how much to invest in the socialization of one's children may be very different from the complex dynamic social processes leading up to an identity group committing an insurrection. There is no specified outcome for an insurrection. We assume the authority sets π in order to avoid an insurrection. Thus, we implicitly assume that the authority must find the insurrection outcome to be worse than being able to reset π, thereby satisfying the constraint. Implicitly, we also assume that the minority might avoid or reduce coercion given a successful insurrection. First, we impose the condition that a larger SSE $q^*(\pi')$ always implies a lower threshold for insurrection. An intuitive reason for this is that a larger minority will, ceteris paribus, have more capability to commit a successful insurrection and, thus, a lower threshold for committing one. Formally, we make the following assumptions regarding the insurrection constraint function:

Assumption 4 Larger non-state identity groups will always have lower insurrection thresholds: The insurrection constraint $\rho(q^*(\pi))$ is a continuous mapping from $q^*(\pi) \in (0, 1)$ to $[0, \pi_{\max}]$. It yields the highest level of coercion for which there is *not* an insurrection. It is monotonically decreasing in $q^*(\pi)$, $\rho'(q^*(\pi)) < 0$ and has a continuous first derivative.

[21]

We insert the endogenous insurrection constraint into (11) to attain the authority's static optimization problem with an endogenous insurrection constraint in the following equation.

$$\max_{\pi} U^{\mathbb{S}} = \min_{\pi} \left[\frac{\Delta u - \pi + C(\pi)}{2\Delta u + C(\pi)} \right] \text{ s. t. } \pi \leq \rho(q^*(\pi_0)) \tag{12}$$

Unless the authority can set π only once, and is unable to subsequently readjust its π, the solution to (12) is not necessarily a DSE. As the insurrection constraint is dependent on $q^*(\pi)$, selecting the optimal $\pi = \pi'$ from an initial condition $q^*(\pi_0)$ may imply that the new insurrection constraint is less binding, $\rho(q^*(\pi_0)) < \rho(q^*(\pi'))$. Hence, the π' solving (12) may be dynamically unstable in the sense that the

[20] The general finding that coercion predicts political violence is well documented; see for instance Choi and Piazza (2016).

[21] An important caveat in applying the model to analysis is to mind the generality of our insurrection constraint. If insurrection capability is sufficiently capital-intensive, the ability to insurrect will not depend on the number of persons holding the non-state identity. Thus, technological and institutional factors will thus be highly important in determining the need for legitimacy to avoid insurrections. We discuss interpretations of the insurrection constraint in "Appendix 3.1".

authority may have an incentive to set a new $\pi'' > \pi'$ in order to attain a lower SSE, $q^*(\pi'') < q^*(\pi')$.

In order to obtain the dynamically stable coercion level an authority *will* implement, we first develop a notion of which coercion levels an authority *can* implement from a given initial condition π_0. We first define the set of sustainable coercion levels, $\mathbf{I}_\Pi \equiv \{\pi : \pi \leq \rho(q^*(\pi))\}$: these are the levels of coercion that, at their corresponding SSE level, do not breach the insurrection constraint. As it is not necessarily the case that all $\pi \in \mathbf{I}_\Pi$ are *implementable* from a given initial condition π_0, there might be no sequence of changes to coercion leading up to an SSE value from which the level can be implemented without making it rational for the minority to insurrect at this level. We formally define the set of implementable coercion levels, $\mathbf{I}_{\pi_0}$, from an initial condition π_0 as described below.

Definition of the set of implementable coercion levels: A coercion level π' is in the set of implementable coercion levels $\mathbf{I}_{\pi_0}$ if and only if there exists a finite sequence $\{\pi_n\}_0^N \equiv \{\pi_0, \pi_1, \pi_2, ..., \pi_N\}$, where $N \in [0, \infty)$ with $\pi^N = \pi'$ that satisfies the following two criteria:

(1) Every coercion level in $\{\pi_n\}_1^N$ is implementable from its previous value: $\pi_n \leq \rho(q^*(\pi_{n-1}))$ for all $n = 1, 2...N$.
(2) Every coercion level in $\{\pi_n\}_1^N$ is sustainable: $\pi_n \in \mathbf{I}_\Pi$ for all $n = 1, 2...N$.

Implementing a coercion level which is not implementable, a breach of condition (1) of the definition of $\mathbf{I}_{\pi_0}$ will immediately lead to an insurrection, as at the time of implementation of a π' it will hold that $\pi' > q^*(\pi_0)$. Implementing a coercion level π', which is implementable, but not sustainable, a breach of condition (2), but not (1), of the definition of $\mathbf{I}_{\pi_0}$, must imply the following aspects: the coercion level imposed is sustainable at the SSE of the initial condition $\pi' < \rho(q^*(\pi_0))$, while it is unsustainable at the SSE corresponding to the imposed level $\pi' > \rho(q^*(\pi'))$. Upon implementing π', there will be no insurrection, but as the non-state identity group grows towards their SSE size, they will at one point choose to insurrect. Thus, the insurrection will occur some time after the implementation of the unsustainable coercion level.

Since we have assumed that both increases and decreases in coercion level can increase the SSE, an implementable but unsustainable coercion level π' can be both higher or lower than the initial coercion level—that is $\pi' > \pi_0$ or $\pi_0 > \pi'$. The intuition of this is straightforward. A reduction in coercion can cause the minority to increase in size. This leads them by assumption to have a greater insurrection capability, thereby making an insurrection more likely to occur even though coercion has been reduced. Whether this occurs is dependent on the change in utility of the insurrection outcome relative to the change in insurrection capability.

This property of the model gives rise to some interesting features of authorities sets of implementable coercion levels $\mathbf{I}_{\pi_0}$: coercion dependence. Authorities with insurrection constraints such that zero is not a sustainable coercion level, $0 \notin \mathbf{I}_\Pi$,

are defined to be strongly coercion dependent: they will be dependent on strictly positive levels of coercion to sustain their state, and impose $\pi > 0$ without any inherent incentive to minimize $q^*(\pi)$. We establish the following as a formal definition of coercion dependence.

Definition of strong, weak, and strictly weak coercion dependence: An authority is defined as *strongly coercion dependent* whenever it cannot impose a zero level of coercion, $\rho(q^*(0)) < 0$, and *weakly coercion dependent* whenever there exist unsustainable levels of coercion π' lower than the initial condition: $\pi' < \pi_0 : \pi' \notin \mathbf{I}_{\pi_0}$. An authority is defined as *strictly weakly coercion dependent* if it is weakly coercion dependent, but not strongly coercion dependent—that is $\pi' < \pi_0 : \pi' \notin \mathbf{I}_{\pi_0}$, while $\rho(q^*(0)) > 0$.

A strongly coercion dependent state implies that the population composition of the state implies that coercion must be use to hold the state together. Therefore, coercion is an inherent part of the state construction. Under strictly weak coercion dependence the non-state identity group prefers to commit an insurrection at some coercion level π' between zero and the current level of coercion, once they attain their equilibrium size at this coercion level, $q^*(\pi')$, but not at the zero level. Thus, the authority will risk an eventual insurrection by setting a coercion level that is too *low*, while being able to sustain the state by a sudden reduction to zero. This implies that a reduction in coercion is possible, but cannot happen gradually without triggering an insurrection. We now analyze how the spread of Protestantism challenged the legitimacy of early-modern European kings and arguably placed them in a situation of coercion dependence.

Coercion dependence in the HRE 1517–1648

The Christian Schism after the Reformation in 1517 and the subsequent spread of the Protestant faith, fuelled by the introduction of the printing press (Rubin 2014) and dismay with the policies of the Catholic Church, led to an increase in religious heterogeneity in early modern Europe. The Haburg elite of the HRE ($\mathbb{S}$) built their legitimacy on the Catholic Church (s), and the introduction of Protestantism (ns) posed a threat to the legitimacy of their state.[22,23]

The initial religious wars and periods of upheaval ended with the admission of religious rights at the Peace of Augsburg (1555). These concessions were made as the rulers realized the unproductiveness of the coercive measures, coupled with an inability to sustain the ensuing military pressure (Wilson 2009). As Johnson

[22] For ease of presentation, we do not distinguish between different Protestant faiths—that is Lutheranism and Calvinism.

[23] The contemporary practices of the Catholic Church itself were considered as illegitimate by members having internalized its identity, most notably one of its own priests, Martin Luther. This created a dissonance within him, prompting him to publicize demands for reform that would spark the creation of a new religious identity, Protestantism. As this identity was introduced into the population, its followers invested heavily in socialization. This lead to a quick spread of Protestantism making the population converges towards a mixed equilibrium as Catholics responded by increasing their levels of socialization, which is in line with the prediction of our baseline model presented in the Appendix.

and Koyama (2013) put it, "This intensified persecution became increasingly ineffective: it served to strengthen the faith of Protestants and encouraged them to organize", the use of coercion proved counter-productive, compatible with a micro-level presence of coercion resentment.

The HRE was not a unified state, but rather a decentralized empire structure of smaller kingdoms with varying degrees of loyalty to the ruling Habsburg family and the HRE authorities. Protestantism served as both a cause of, and an excuse for, peripheral resistance against central authorities. Lower-level princes actively used religious cleavages and changed their religious affiliations to challenge the hegemony of the Emperor, build alliances, and gain influence (North and Thomas 1973). This demonstrates how religious homogeneity was a necessity for maintaining a strong state, and why implementing the identity of the state was considered imperative for keeping the Empire united and under the control of the ruling Habsburg elite.

After granting Protestants (*ns* identity) the right to practice their faith at the Peace of Augsburg, Emperor Charles V (1516–1556) still considered Protestants as a challenge to his powers and, at his death, the Habsburg family was divided between moderate and traditionalist views of which policies should be adopted towards the Protestants (Wilson 2009; Nexon 2009). The moderates wanted to pursue a non-confrontational line and build legitimacy for both faiths, while the traditionalists wished to purge the empire of Protestantism through the use of force—that is to increase π. The Habsburgs recognized that the current coercion level was in the inefficient $(\underline{\pi_q}, \underline{\pi^e_q})$ interval, but were uncertain and divided on the direction forward, and whether coercion levels beyond $\underline{\pi^e_q}$ would trigger an insurrection and whether the current level of coercion was sustainable (Wilson 2009).

In 1618, it became evident that Ferdinand II, who had pursued strong anti-Protestant policies in Austria, would be the successor to the throne. This further increased tension in the Protestant-dominated region of Bohemia. In the period from the Peace of Augsburg in 1555 to 1618, there was an increase in the number of Protestants (Cantoni 2015). This is compatible with an increased investment in socialization and consequently, growth in the prevalence of q in line with a micro-level coercion resentment. The imminent coronation greatly increased resentment towards the Empire and the anticipated change towards a more confrontational policy. Through the lens of the model, this can be seen as tipping the insurrection constraint following resentment towards the emperors' new-found ambition for a counter-reformation. It clearly acted as a prerequisite for the approaching conflict. The renewed program of confrontational religious homogenization that was anticipated after the coronation of the more religiously dedicated Emperor Ferdinand II, strongly contributed to the Protestants' insurrection at the Second Defenestration of Prague, sparking the Thirty Years' War (1618–1648) between Protestants and Catholics in Germany (Wilson 2009).

The coercive policies of the Habsburgs produced armed conflict and did not lead to religious homogeneity in the HRE. The coercive policies did not in any sense produce a stable long-term equilibrium. The French Crown had faced the similar development of Protestants, as the HRE would yet remain a unified and

religiously homogenous country. We now consider the possible long-term equilibriums, DSE of the model, and how they align with the coercive policies of the French experience of the Reformation.

2.3 Dynamically stable equilibrium

We now find the authorities' optimal level of coercion amongst the implementable coercion levels, $\mathbf{I}_{\pi_0}$, by developing the notion of a DSE, π^{DSE}.

Definition of a DSE: A DSE is defined as a coercion level and an SSE $\{\pi^{\text{DSE}}, q^*(\pi^{\text{DSE}})\}$, such that π^{DSE} is the optimal coercion level if π^{DSE} is equal to the initial condition π_0, $\pi_0 = \pi^{\text{DSE}}$.

We find π^{DSE} by solving the following equation.

$$\pi^{\text{DSE}} \equiv \left\{ \underset{\pi}{\text{argmax}} U^{\mathbb{S}} = \min_{\pi \in \mathbf{I}_{\pi_0}} \left[\frac{\Delta u - \pi + C(\pi)}{2\Delta u + C(\pi)} \right] \right\} \tag{13}$$

Any authority that can infinitely reset π will always be at, or in a sequence $\{\pi_n\}_0^{N-1}$ leading up to, a DSE $\{\pi^{\text{DSE}}, q^*(\pi^{\text{DSE}})\}$.[24] We define a *strategic* constraint, if the coercion level at the bound of $\mathbf{I}_{\pi_0}$ as a fix-point of the insurrection constraint, π_{fix}: this will be the highest level of coercion implementable where a non-state identity does not commit an insurrection. We note that any DSE coercion level will either be unconstrained, equal to a strategic constraint, or at the feasibility constraint at the end of the $[0, \pi_{\max}]$ interval, or, a constrained level at an internal minimum of $q^*(\pi)$. We establish this as Lemma 4.

Lemma 4 Dynamically stable equilibria: *For any initial condition π_0, the DSE π^{DSE} is a coercion level equal to either:*

I. *the unconfrontational level of coercion as an interior point of* $\mathbf{I}_{\pi_0}$: $\pi^{\text{DSE}} = \pi_{\underline{q}}$
II. *a strategic constraint at the upper bound of* $\mathbf{I}_{\pi_0}$: $\pi^{\text{DSE}} = \pi_{\text{fix}}$
III. *the upper feasibility constraint at the bound of* $\mathbf{I}_{\Pi}$: $\pi^{\text{DSE}} = \pi_{max}$.

When $\{\Delta u, C(\pi)\}$ is sufficiently low, such that $q^*(\pi)$ is of class III), π^{DSE} will either be equal to a strategic constraint $\pi_{\text{fix}} \in [0, \pi_{\underline{q}})$ or the lowest π' attaining $q^*(\pi) = 0$, depending on whether the insurrection constraint function is such that

[24] The DSE π^{DSE} will be unique corresponding to every feasible π_0. The monotone derivative of the insurrection constraint function, $\rho'(q^*(\pi)) < 0$, implies that at any iteration, there can never be another $q^*(\pi)$ that yields a higher insurrection constraint than the lowest attainable $q^*(\pi)$. Hence, there cannot be a lower reachable $q^*(\pi')$ than the π' reachable through minimizing $q^*(\pi)$ in every iteration. In other words, maximizing capability to reach any long-term goal and maximizing short-term gains will imply equal behaviour. The prediction of the DSE is robust to the introduction of time preferences when coercion is costless and $\rho(q^*(\pi))$ is monotonically decreasing in $q^*(\pi)$.

π' is implementable from π_0. For $q^*(\pi)$ of class I), the dynamically stable π^{DSE} is equal to $\pi_{\underline{q}}$ if this is implementable, and equal to some strategic constraint $\pi_{\text{fix}} > \pi_{\underline{q}}$ if not. For $q^*(\pi)$ of class II), authorities will end up in stable gunpoint equilibria with two identity populations when $0 < q^*(\pi_{\max})$ whenever $\pi_{\max}$ is implementable. If there exists an implementable π', such that $q^*(\pi') = 0$, then the population will approach single identity equilibrium. If this is not the case, then either π^{DSE} is equal to $\pi_{\underline{q}}$, the unconfrontational level of coercion, or the equilibrium must be a strategic constraint, either at some coercion level above $\pi_{\text{fix}} \in [0, \pi_{\underline{q}})$ or below $\pi_{\text{fix}} \in (\pi^e_{\underline{q}}, \pi_{\max})$, the open interval of inefficient coercion levels, $(\pi_{\underline{q}}, \pi^e_{\underline{q}})$.

The equilibrium $\pi^{\text{DSE}} = \pi_{\underline{q}}$ is the only one where the authority restrains its use of coercion in equilibrium. Here, the authority imposes a coercion level strictly lower than the highest implementable coercion level. The equilibrium $\pi^{\text{DSE}} = \pi_{\text{fix}}$ is given by a fix-point of $\rho(q^*(\pi')) = \pi'$ and implies a coercion level at a binding insurrection constraint. Thus, the coercion level will be at the highest level where it does not trigger an insurrection. Finally, the equilibrium where $\pi^{\text{DSE}} = \pi_{max}$ can be understood as an equivalent of legitimacy at the "barrel of a gun". The gunpoint level of legitimacy is defined as the legitimacy that can be achieved at the $q^*(\pi)$ corresponding to socialization investment at its SSE value of $\pi_{\max}$.

Equilibriums of the model and the French Huguenots

Similar to the wars of religion in the HRE, the French Crown had waged war with its Protestant population, the Huguenots, from the beginning of the Reformation (1517). Recognizing the unproductiveness of its policies during the French Wars of Religion, the French Crown settled for a non-confrontational equilibrium with the Edict of January at St. Germaine in 1562. Protestantism was decriminalized, but the Huguenots were not allowed to worship publicly—an illustrative example of a non-confrontational level of coercion, $\pi_{\underline{q}}$, in our model.

Prior to the decision to once more outlaw Protestantism with the Revocation of the Edict of Nantes in 1686, advisors close to the French king, Louis XIV, recognized the potential counterproductiveness of this policy (Sutherland 1988). Initially, coercive measures were introduced gradually, as French historian Elisabeth Labrousse puts it: "measures, therefore, had to be constantly presented, albeit with a good deal of sophistry, not as aggressive sanctions but simply as a withdrawal of the kings' favors from the minority." (Sutherland 1988). Marginal changes to the coercion level were gradually imposed to reduce salience and potential counter-reactions. Louis XIV's advisors recommended a continuation of this policy by identifying the Huguenots as schismatic, a measure more gentle than outlawing Protestantism. However, Louis XIV chose the stricter, more confrontational line and the death penalty for Protestants was introduced in France on July 1, 1686. While it appears that the king's advisors recognized the potential for a reaction to his policies, the king was surprised by the negative response and mass exodus. His hopes had been for reformation rather than relocation as a Protestant response.

While the granting of religious rights in 1547 to the Huguenots was given in order to make peace with a politically and militarily powerful group, the revocation of the rights was made to a small group that posed little or no military threat following a long and gradual increase of coercive measures by the French Crown (Rae 2002). Thus, in terms of the model, because $\pi_{\underline{q}}$ was not the upper limit of the set of implementable coercion levels, the crown did not hit any strategic constraints along the program of increases in π towards the DSE, which would turn out to be the gunpoint equilibrium. In line with the functional form established in Lemma 2, once the coercion levels approached an inefficient level, Louis XIV went directly to a clear case of a gunpoint threat and avoided any potentially inefficient coercion levels in $(\pi_{\underline{q}}, \pi^e_{\underline{q}})$.

The revocation happened against the backdrop of a conflict with the largely Protestant Dutch and dismay among the Catholic of the relative economic success of the Huguenots minority (Thompson 1908). The policy held no military cost for the French king but had a reputational cost. The reactions from foreign kings were negative, condemning the treatment of the French Huguenots (Labrousse 1985 in Rae 2002), perhaps indicating the nascent expectations of minority rights being respected in international relations. The view of the kings' motivation, as expressed by Labrouse, was "the Revocation attempted to abolish a religious heresy because it was thought to harbour a political heresy".[25]

2.4 Path dependency

We have now developed a formal relation from the coercion decisions of the authority, to the prevalence of the minority identity, to their ability to insurrect, which constraints the authority's ability to coerce. This dynamic relationship is captured by the shape of the insurrection function illustrated in Fig. 4.

Assume two different authorities face different initial conditions, $\overline{\pi_0} > \underline{\pi_0}$. The authorities have the same set of preferences in their population Δu, the same response to coercion $C(\pi)$, and have equal available military technology $\rho(\cdot)$. Then, they face the same insurrection function $\rho(q^*(\pi))$ and have the same set of sustainable coercion levels $\mathbf{I}_{\Pi}$. Assume that both non-state identity groups would choose to lower their socialization rather than insurrect, given a sufficiently small size, $q^*(\pi')$. However, this size, $q^*(\pi')$ is not attainable by gradual increases from the initial conditions of one of the authorities but not the other authority—that is $\pi' \in \mathbf{I}_{\overline{\pi_0}}$ while $\pi' \notin \mathbf{I}_{\underline{\pi_0}}$. Then the equilibrium size of the non-state identity is determined by some pre-history creating the differences in which coercion levels can be implemented. A

[25] Scholars studying the period surmise that without the actions of the state, the Huguenot identity might have withered away in the absence of persecution (Labrousse 1985 in Rae 2002); thus, the policy could be viewed from a purely realist perspective as an error. The identification of emigrated French Protestants as Huguenots, a separate identity from the Catholic French, would remain strong, albeit outside France (Sutherland 1988). This insight is interesting in the light of the model. It indicates that either the "cultural memory" of persecution, $C(\pi)$, as having a long-term identity building effect or a persistent high investment in socialization.

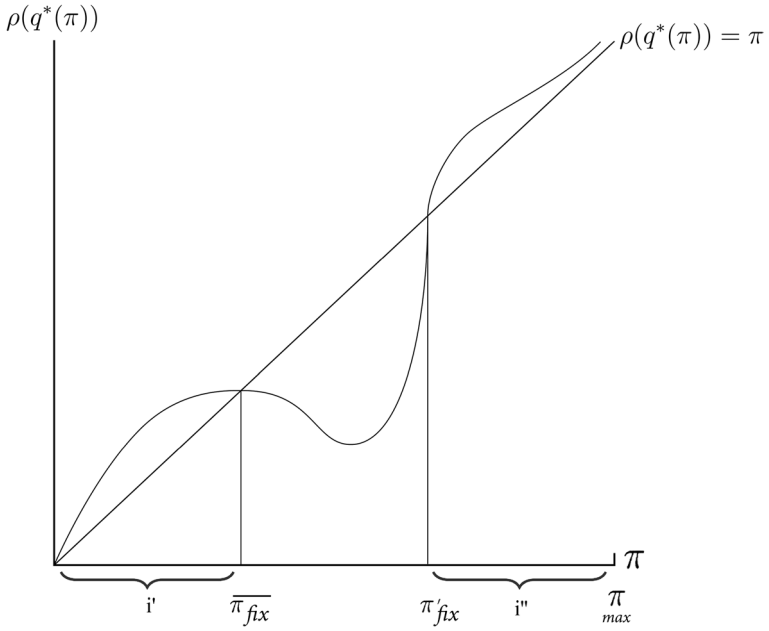

Fig. 4 The relationship between the highest implementable coercion level $\rho(q^*(\pi))$ and current coercion level π. *Note:* The curved line is an example of the insurrection constraint function, $\rho(q^*(\pi))$, while the 45° degree line is the fix-point-line. Any $\pi_0 \leq \overline{\pi}_{\text{fix}}$ yields a set of implementable coercion levels containing one subset $\mathbf{i}'$, $\mathbf{I}_{\pi_0} = \mathbf{i}' = [0, \overline{\pi}_{\text{fix}}]$, while a $\pi_0 \geq \pi'_{\text{fix}}$ yields a set containing two subsets $\mathbf{i}'$ and $\mathbf{i}''$: $\mathbf{I}_{\pi_0} = \{\mathbf{i}', \mathbf{i}''\} = \{[0, \overline{\pi}_{\text{fix}}], [\pi'_{\text{fix}}, \pi_{\max}]\}$

prime example of what can create different pre-history is borders drawn up to determine the final sizes of non-state identity groups, which often happens according to a number of factors external to the model. It is then pre-history and not technology or differences in preferences that determine insurrection capability and, thus, DSE prevalence of the non-state identity group.

Formally, when an insurrection function is such that different initial conditions, π_0, generate different $\mathbf{I}_{\pi_0}$ sets, the model will have path dependency. Different initial conditions will yield different DSE, π^{DSE}.[26] We conclude our formal analysis by developing path dependency as a lemma

Theorem of path dependence DSE is inherently given by the history of the polity if and only if there exist coercion levels that are sustainable but unimplementable from certain initial conditions.

[26] For an insurrection function $\rho(q^*(\pi))$ with two crossings at the fix-point-line, such as the one described in Fig. 1, $\mathbf{I}_\Pi$ will consist of two disjoint subsets, $\mathbf{i}$ — one lower $\mathbf{i}'$ and one upper $\mathbf{i}''$. Assume that $\mathbf{I}_{\overline{\pi_0}}$ is equal to $\{\mathbf{i}'', \mathbf{i}'\}$; then, $\pi_0 \in \mathbf{i}'$ and $\overline{\pi_0} \in \mathbf{i}''$ will produce different sets of implementable coercion levels $\mathbf{I}_{\overline{\pi}_0} \neq \mathbf{I}_{\pi_0}$, depending on whether there is any way of implementing the minimal π' in the upper subset $\mathbf{i}''$ from the lower $\mathbf{i}'$—that is, if $\rho(q^*(\inf \mathbf{i}')) < \inf \mathbf{i}''$.

Note that, whenever multiple dynamically stable equilibria exist, an exogenous temporary increase in insurrection capability can move the DSE between sets of long-term implementable coercion levels. Consequently, the model makes the prediction that coercion levels and corresponding SSE prevalence of identity groups will, in certain cases, be inherently dependent on history in conjunction with the included long-term equilibrating factors.

This concludes our formal analysis. Most results should naturally generalize for other nonlinear functional forms where the marginal effectiveness of coercion varies with its level—that is, dynamically stable equilibria will either be at a π_{fix} or at a local or global minimum of $q^*(\pi)$ within $[0, \pi_{\max}]$. For example, assume an S-like coercion resentment function tantamount to the one assumed with several consecutive convex and concave areas. The results of the analysis would naturally generalize to this functional form. For each sufficiently concave interval there would be an additional inefficient interval of coercion and for each sufficiently convex interval there could be an interior stable equilibrium.

Path dependency and a comparative perspective on the HRE and France: The role of the Peace of Westphalia

The aftermath of the Thirty Years' War, the Peace of Westphalia, has been thoroughly studied in terms of international relations and considered to mark the beginning of the modern state system. Interpreting the new institutional paradigm of international relations in Europe as relaxing the insurrection constraint, $\rho(q^*(\pi))$, because of a lower risk of foreign involvement, can account for the freezing pattern in the map of religious identities after 1648 in Europe, which is thoroughly documented in the political science literature (Rae 2002; Nexon 2009; Tilly and Ardant 1975; Rokkan 1999).

The human suffering in the war estimates of population loss range from 10 to 45% (Theibault 1997) within the borders of modern-day Germany increased both demands for a new paradigm and respect for new institutional rules of international relations in Europe. It also left Germany a country religiously divided between Protestants and Catholics. Furthermore, among the elites, the Thirty Years' War was seen as an example of how not to wage war, an example of the dangers of religious passions and mercenary armies (Philpott 2001).

Among the changes agreed at Westphalia was the principle of territoriality which created at least a minimal requirement for a legitimate claim to the territory. It tied religious identities to territorial identities, thereby increasing the need for religious homogeneity. The treaty obliged the king to have the same religious affiliation as that of his polity (Wilson 2009), thereby reducing the incentives of changing faith to gain power. Furthermore, it outlawed the use of religious tension in countries as a legitimate reason for engagement in civil wars. Together with the further delegitimization of mercenary armies, these measures effectively decreased the insurrection constraint. Religious minorities reduced their ability to convert economic resources to military capability and their ability to further the interests of foreign powers as strategic "jus ad bellum" as a pretext to go to war.

The model points to how lower insurrection constraints will lead to the lower prevalence of non-state identity either through a quicker convergence towards

equilibrium or by enabling the authority to impose a program towards the gunpoint equilibrium. Hence, the model can account for how the Peace of Westphalia increased internal homogenization as a consequence of the delegitimization of using religious schisms as a pretext for foreign involvement in internal conflicts.

While the attempts to homogenize the HRE lead to insurrection, foreign involvement, and subsequently a religious division of the Empire, the potential Huguenot mobilization could not be turned into a pretext for foreign involvement and a consequent military threat to the French king under the new institutional framework. This absence of a threat from neighboring countries greatly relaxed the insurrection constraints as governments could focus on internal enemies when pursuing homogenization, thereby predicting closer alignment between territory and state identities (Nexon 2009).

The changing military technology, away from professionalized soldiers with training in the use of both firearms and swords, towards mass armies primarily dependent on gunpowder, placed a higher military value on draftable citizens. In the language of the model, changing military technology leads to a higher $\rho'(q^*(\pi))$. The insurrection constraint became more sensitive to the prevalence of the state identity as military capabilities became more sensitive to mass support. This, in turn, lead to an increase in the demand for homogenization of populations, thereby enabling the draft large standing armies to stand against external threats, which would propel the development of consolidated states.

3 Concluding remarks

We now return to the two stylized facts we set out to explain, variation in coercion use and variation in the persistence of non-state identities. To explain coercion use, we have developed a model that demonstrates how the micro-foundations of coercion resentment can be used to understand legitimacy-maximizing authorities. This model is built around the premise that authorities recognize coercion resentment and adopt their policies accordingly. We argued that the assessed monarchs of Early Modernity and Stalin restrained their use of coercion in response to strategic constraints in a manner that is explainable by our theoretical framework. Then, we extended the model to account for how the size of the non-state group affects their ability to insurrect. Thus, their size affects constrains on the use of coercion. Formally, we have been able to establish what constitutes long-term stable equilibrium and showed that this understanding might help explain coercion use against French Huguenots. Once the prevalence of a non-state group reduces, their ability to insurrect reduces. Thus, previous coercion levels might become an out-of equilibrium outcome. Further, we developed a definition of coercion dependence to show why coercion can be the only way some authorities could sustain their rule, as lowering coercion levels in places like the HRE could lead to internal instability. Finally, we formalized a notion of path-dependency that could help explain the freezing patterns of religious and national identities in Early-Modern Europe. Our proposed mechanism is that once non-state identities that were sufficiently large had an ability to

insurrect, their presence became an long-term equilibrium part of the state in which they resided.

There are several potential extensions of this model that can address related questions in future research.

Parental preferences for coercion An applied problem is considering that the state identity parents can fully or partially set the coercion level. The state identity parents effort in socialization decreases in the levels of state coercion of the *ns* identity.[27] Assuming that the majority of parents do not have utility in the outcomes of the state, parents of state identity will prefer the coercion level that balances the trade-off between private preferences for lowered socialization and social preferences for future generations of state and non-state identity children and parents. Exploring a model where parents can set coercion levels in conjunction with historical evidence can shed light on processes where democracies become coercive or authoritarian. Further theoretical work along these lines can address the question of the extent to which totalitarian policies emerge from political demand or political supply.

Evolutionary properties of state competition Assume that authorities are naïve regarding the effect of coercion. If so, authorities that impose coercion levels within the set of implementable coercion levels will endure, and others will perish from insurrections. Future theoretical analyses that apply the set of implementable coercion levels can tie together empirical evidence of historic and pre-historic processes of state competition in new ways.[28]

Costly coercion under discounting Exogenous variations in insurrection costs, variations in the benefits of legitimacy of the authority, and variations in seceding for the minority can arise from factors such as rough terrain or rents from natural resources. Hence, there are reasons to assume that the set of implementable coercion levels might be different for authorities with access to the same military technology, and that authorities might choose to impose different coercion levels because of differences in benefits, costs, of legitimacy.[29] Furthermore, under costly coercion,

[27] This holds as long as the state identity is also the majority identity. If the state identity is a minority, the issue depends on functional form—that is, socialization responses to the use of coercion.

[28] For example, consider an extension of this model where populations of polities of uniform size and initial conditions compete. Assume that the authorities are naïve regarding the effect of coercion but are able to use military capabilities externally to overtake neighbouring states. The population of polities in such a model presumably will, over time, converge towards only polities that impose the dynamically stable equilibria. Room for deviation from optimal polices will grow with differences in the relative sizes of polities and their initial conditions. Hence, it appears that the proposed equilibria can arise from state competition, in line with the arguments set forth in Tilly (1992), even under conditions of naïveté regarding the effects of, and constraints on, coercion. Note, however, that for both this argument and the argument put forth in Tilly (1992), the actual solution might be more complex than first expected.

[29] An extended model that includes these properties could provide a micro-foundation to Barfield (2010)'s explanation of the high ethnolinguistic variance in Afghanistan. He places emphasis on how rough terrain, yielding a low insurgency cost compared with the low value of attaining legitimacy, together with multiple historic influences—that is multiple seed identities, and low benefits of having legitimacy, have contributed to the large cultural heterogeneity observed in Afghanistan.

equilibrium outcomes will also be determined by the time preferences of the authority. There will be a trade-off between the discounted future benefit of legitimacy and the present cost of coercion. This could account for why different dynastic structures — that is, with more or less direct hereditariness of power, could lead to different policies. In modern democracies, such differences in incorporating the future can arise from variations between candidate politics versus party politics.

Framing and timing of coercion It appears likely that effects such as cultural memory, incentives of community leaders, and sluggishness in investment in military technology, change the effect of coercion and, consequently, the set of implementable coercion levels. Furthermore, different programs in terms of how gradual changes are and how they can be framed, will imply that the set of reachable coercion levels will differ for different strategies and different pre-histories.

The framework's explicit modelling of population responses, together with the possibility of strategic analysis, makes it a potential tool for policy analysis for an external agency constraining an authority's use of coercion. Generally, limiting the level of accepted coercion will depend on the views of external agencies regarding the ratio of the cost of commission versus the cost of omission—that is the cost of limiting coercion and the benefit of limiting the suffering caused by coercion itself. Further research can develop a theory that incorporates ethnic compositions and power relations as inputs to predict the initial states—that is polity borders that can create sustainable uncoercive states, the cost of reaching these states, and where the pitfalls of state failure lie.

Investigating general models of these dynamics play an important role in using the conflicts of the past to avoid conflicts in the future. Further, we can seek to understand how the diversity of identities can be an equilibrium outcome in the face of legitimacy-maximizing authorities. Although technology, beliefs, and institutions change, as long as human nature remains stable, the past will be informative of the future.

Acknowledgements Open Access funding provided by Norwegian School Of Economics. The project was financed by support from the Research Council of Norway through its Centres of Excellence Scheme, FAIR Project No. 262675 and research Grant 185831, and administered by FAIR-The Choice Lab, NHH Norwegian School of Economics and the Japan Society for the Promotion of Science. I would like to thank Avner Greif, Chiaki Morgiguchi, Bertil Tungodden, Timur Kuran, Claude Diebolt, seminar participants at The 2018 Cultural Evolution Society Conference in Arizona, Hitotsubashi University Economic History Seminar Series and two anonymous referees for all their support on this version. I would also like to thank Robert Arons, Rodney Beard, Elias Braunfels, Gary Charness, Erik Eikeland, Jon Fiva, Armando Jose Garcia Pires, Peter Hatlebakk, Ola Honningdal Grytten, Rune Jansen Hagen, Thor Øivind Jensen, Jo Thori Lind, Jared Rubin, Daniel Spiro, Ragnar Torvik, Simen Ulsaker, Tom Grimstvedt Meling, Moti Michaeli, Linda Nøstbakken, Frederik Willumsen, ASREC 2017 Sixteenth Annual Conference in Boston, the 5th Annual Graduate Student Workshop at IRES at Chapman University, the 12th Nordic Conference in Development Economics, the 2015 Meeting of the Norwegian Association for Economists, CMI, The 2015 UIB-NHH Ph.D. Workshop in Economics, ESOP Seminar at UIO, NMBU Ås Economics Brown Bag Seminar, LEMO Seminar at NHH, UIB System Dynamics Seminar and UIB Philosophy Seminar in Political Theory for insightful comments and Anne Liv Scarce for research assistance.

Appendix 1: Bisin and Verdier (2000, 2001) model of socialization

Following Bisin and Verdier (2000, 2001), we introduce an overlapping generations model where parents invest in costly socialization to make their child internalize the identity of the parents. First, the basics of the model and the mechanisms of socialization are developed. All the results here mirror the results from Bisin and Verdier (2001). We then develop assumptions regarding the parents' utility function and derive its implications.

The model

The population consists of a continuum of agents who live in two periods, as a child at time t and as a parent at time $t+1$. Each agent produces one offspring. Thus, the size of the population remains stable. There are two identities, $m \in \{a, b\}$. identities are mutually exclusive. A portion q_t of the parent population holds identity ns at time t, while $1 - q_t$ holds identity s. Identities are transmitted from one generation to the next through parental socialization from parent to child, or through oblique transmission, the influence of the general population. The probability is τ^m that parental socialization is successful and the child adopts the identity of the parent, and $1 - \tau^m$ that parental socialization fails. If parental socialization fails the child is obliquely socialized and the offspring will adopt either identity ns or identity s with a probability equal to the identity's prevalence in the population. A child who internalizes identity m is referred to as an m identity child. Let P^{mn} be the probability that an individual of identity m has an n identity offspring.

$$P^{aa} = \tau^{ns} + (1 - \tau^{ns})q_t \,,\; P^{ab} = (1 - \tau^{ns})(1 - q_t) \tag{14}$$

$$P^{bb} = \tau^{s} + (1 - \tau^{s})(1 - q_t) \,,\; P^{ba} = (1 - \tau^{s})q_t \tag{15}$$

The portion of the population with identity a at time $t+1$, q_{t+1}, is then given as follows.

$$q_{t+1} = q_t P^{aa} + (1 - q_t)P^{ba} = q_t + q_t(1 - q_t)(\tau^{ns} - \tau^{s}) \tag{16}$$

From (16), it follows that the change in the share of identity ns individuals is given by $q_t(1 - q_t)(\tau^{ns} - \tau^{s})$: the difference in the probability of successful parental socialization times the product of the share of identities.

Parents choose τ^m to maximize expected utility by balancing the cost of parental socialization, denoted by the function $H(\tau^m)$, and the benefit of a higher probability of successful parental socialization. Let the utility of an m identity parent having an n identity child be denoted u_n^m, then using (14) and (2), we attain the following utility function U^m for parents.

$$U^{ns} = [\tau^{ns} + (1-\tau^a)q_t]u_a^{ns} + (1-\tau^{ns})(1-q_t)u_a^s - H(\tau^{ns}) \tag{17}$$

$$U^s = [\tau^s + (1-\tau^b)(1-q_t)]u_b^s + (1-\tau^s)q_t u_b^{ns} - H(\tau^s) \tag{18}$$

We now impose some assumptions on the parents' preferences for their child's identity and the cost function of parental socialization. First, we assume that parents prefer their child to have the parents' identity.

Assumption A1 Own identity preference Parents prefer their child to have the same identity as themselves: $u_a^{ns} - u_b^{ns} > 0,\ u_b^s - u_a^s > 0$.

Second, the utility loss of having a child internalize an opposing identity is assumed to be symmetric for the two types of parents. Defining $\overline{u}$ as the utility derived from the child having the parents' own identity and $\underline{u}$ as the utility derived from having the opposing identity, we can write the following assumption.

Assumption A2 Symmetric utility loss of opposing identity Parents of ns and s identity have symmetric utility loss in having children of opposing identity: $u_a^{ns} - u_b^{ns} = u_b^s - u_a^s = \overline{u} - \underline{u} = \Delta u$.

Third, we assume that the cost of socializing the child into the preferred identity $H(\tau^m)$ obeys the Inada conditions.

Assumption A3 Inada assumptions Inada conditions apply to the cost of investment in parental socialization: $H'(\tau^m) \geq 0, H'(0) = 0, \lim_{\tau^m \to 1} H'(\tau^m) = \infty, H''(\tau^m) > 0$.

The first part of Assumption A3 states that the marginal cost increases with the probability of success, and the second that there is no marginal increase in the cost of socialization at no parental socialization, $\tau^m = 0$. The third and fourth parts of Assumption A3 state that the marginal cost approaches infinity as the probability of having a child successfully socialized into the preferred identity approaches certainty, and that the increase in marginal cost is strictly increasing in τ^m. The assumption of no increase in cost at $\tau^m = 0$ implies that τ^m will be strictly positive whenever the utility of having successful parental socialization is strictly positive for m identity parents. The assumption that the cost of socialization grows towards infinity implies there will always be some failed parental socialization leading to oblique socialization. Hence, there will always be some children obliquely socialized into the opposing identity in mixed identity populations. We can now derive the optimal levels of τ^m from (17) and (5), which are given by the first order conditions (FOCs).

$$H'(\tau^{ns}) = (1 - q_t)\Delta u,\ H'(\tau^{s}) = q_t \Delta u \tag{19}$$

The optimal level is given by the expected marginal benefit of investing in parental socialization, being equal to the marginal cost. From Assumption 3, the Inada conditions and (19), we can establish the following lemma.

Lemma A1 *The smallest identity group always invests more in parental socialization: $\tau_t^s \leq \tau_t^{ns}$ if and only if $q_t \leq (1 - q_t)$.*

As the benefit of having a child with the parents' identity is assumed to be symmetric, a difference in investment must imply a difference in the cost of failed parental socialization. Any difference in the utility of failed socialization arises, as the probability of the child obliquely internalizing the preferred identity differs because of different group size. Minority parents have a higher probability of their child internalizing the majority identity obliquely if parental socialization fails, and consequently invest more in socialization, hence Lemma 1.

A steady state equilibrium (SSE) level of q, denoted as q^*, is reached when $q_t = q_{t+1}$. It follows from (16) that for $q_t = q_{t+1}$ to be fulfilled, $q_t(1 - q_t)(\tau^{ns} - \tau^s) = 0$ must hold. This is the case for $q_t = q_{t+1} = 0$, $q_t = q_{t+1} = 1$, i.e. single identity populations, or, as will be shown, at the interior SSE where $\tau^a = \tau^b$. In cases of $q^* = 0$ or $q^* = 1$, there will be no utility gain from parental socialization as all individuals in the population will have the same identity, and oblique socialization will lead to the preferred identity of the parent. The single identity equilibrium is, however, unstable in the event of external shocks; if one parent of another identity enters the population, this parent would choose a very high investment in parental socialization because the probability of the child adopting the desired identity in the case of oblique socialization would be very low. This would be repeated for future generations and consequently, the prevalence of introduced identity of the minority would grow until the unique interior $q^* = \frac{1}{2}$ is reached.

Lemma A2 *There is a unique stable interior SSE at $q^* = \frac{1}{2}$.*

The only stable equilibrium is $q^* = \frac{1}{2}$; any initial population with a q different from one or zero will converge towards it. For the SSE, the share of minority identity individuals will grow with time as the smaller identity group invests more in socialization, as stated by Lemma 1, until again $q_t = q_{t+1} = \frac{1}{2}$. The fact that the stable interior is $q^* = \frac{1}{2}$ arises because of Assumption 2: symmetry of preferences. Asymmetrical preferences where an interior SSE exists at $\tau^{ns} = \tau^s$, lead to an asymmetrical, i.e. $q^* \neq \frac{1}{2}$, stable SSE.[30]

[30] The assumption of symmetric preferences is made in order to focus on the role of the state rather than on any difference between the desirability of the identities themselves. The following analysis generalizes for asymmetrical preferences.

Proofs for baseline model

Proof of Lemma A1

Lemma A1: The smallest group always invests more in parental socialization: $\tau_t^s \leq \tau_t^{ns}$ if and only if $q_t \leq (1 - q_t)$.

Proof Suppose $q_t < (1 - q_t)$, it then follows from (19) that $H'(\tau_t^s) < H'(\tau_t^{ns})$. By the Inada condition of $H''(\tau^m) \geq 0$ in Assumption 3, it follows that $\tau_t^s < \tau_t^{ns}$. The only if part follows from the fact that there are only two identities. □

Proof of Lemma A2

Lemma A2: There is a unique stable interior SSE at $q^* = \frac{1}{2}$.

Proof We show that for (*i*) there exists a unique interior SSE at $q^* = \frac{1}{2}$, (*ii*) and that it is stable.

(i) Existence and uniqueness of an interior $q^* = \frac{1}{2}$.

An SSE level of q, denoted q^*, is reached when $q_t = q_{t+1}$. It follows from (16), $q_{t+1} = q_t + q_t(1 - q_t)(\tau^{ns} - \tau^s)$ that for $q_t = q_{t+1}$ to be fulfilled, $q_t(1 - q_t)(\tau^{ns} - \tau^s) = 0$ must hold. Hence, at any interior SSE, i.e. $q^* \in (0, 1)$, $\tau^{ns} = \tau^s$. From (19), it follows that this implies $q_t = (1 - q_t)$, which gives $q^* = \frac{1}{2}$.

(ii) Stability of $q^* = \frac{1}{2}$.

We will show that for any $q_t \in (0, 1) \neq q^*$ there will be convergence towards q^*. Suppose $q_t > q^*$, it then follows from (19) that $H'(\tau_t^s) < H'(\tau_t^{ns})$. By Lemma 1 it follows that $\tau_t^s > \tau_t^{ns}$. By (16), $q_{t+1} < q_t$ when $\tau_t^s < \tau_t^{ns}$ and $q_{t+1} < q_t$ for $\tau_t^s > \tau_t^{ns}$. Thus, any $q_t \in (0, 1)$ will converge to q^*. In other words, (0, 1) is a q^* basin of attraction.

□

Appendix 2: Proofs

Proof of Lemma 1

Lemma 1: For all pairs of $\{\pi, \Delta u\}$, two exterior SSEs exist. For some, but not all, pairs of $\{\pi, \Delta u\}$ a unique stable interior SSE exists, given by $q^*(\pi) = \frac{\Delta u - \pi + C(\pi)}{2\Delta u + C(\pi)}$. Imposing a coercion level π' corresponding to an interior SSE, $q^*(\pi') \in (0, 1)$, from an initial interior SSE, $q^*(\pi_0)$, will make q converge to $q^*(\pi')$.

Proof *(i) For all pairs of* $\{\pi, \Delta u\}$, *two exterior SSEs exist.*

By definition, a SSE is given by $q_t = q_{t+1}$. For $q_t \in \{0, 1\}$, (16) implies that $q_t = q_{t+1}$ for any pair of $\{\pi, \Delta u\}$.

(ii) For some, but not all, pairs of $\{\pi, \Delta u\}$, *a unique stable interior SSE exists, given by* $q^*(\pi) = \frac{\Delta u - \pi + C(\pi)}{2\Delta u + C(\pi)}$.

Suppose $\frac{\Delta u'-\pi'+C(\pi')}{2\Delta u'+C(\pi')} = \frac{1}{2}$ and that $\{\pi', \Delta u'\}$ is the imposed π and Δu. We now show that this implies there exists an SSE where $q^*(\pi') = \frac{1}{2}$.

Consider $q_t = \frac{1}{2}$. As $\Delta u' - \pi' + C(\pi') = 1 > 0$, (4) implies that $\tau_a > 0$ for $q_t > 0$. As $\Delta u' + \pi' > 0$, we see from (5) that $\tau_b > 0$ for $(1 - q_t) > 0$. For $q_t = \frac{1}{2}$ to be an SSE, it follows from (16) that $\tau_a = \tau_b$. This implies the left side of (4) should equal the left side of (5). Under $q_t = \frac{1}{2}$, this gives $\frac{1}{2}(\Delta u - \pi' + C(\pi')) = \frac{1}{2}(\Delta u + \pi')$. This implies $2\pi' = C(\pi')$, which is consistent with $\frac{\Delta u'-\pi'+C(\pi')}{2\Delta u'+C(\pi')} = \frac{1}{2}$.[31]

Uniqueness of the interior SSE $q^*(\pi)$ is trivially given by the fact that $q^*(\pi) = \frac{\Delta u-\pi+C(\pi)}{2\Delta u+C(\pi)}$ is a single-valued function. The equilibrium is stable as (0, 1) is a $q^*(\pi)$ basin of attraction following the lines of the argument in the proof of Lemma A2 part *(ii)*.

We finally show that for some $\{\pi'', \Delta u''\}$, no interior SSE exists. Suppose $\{\pi'', \Delta u''\}$ is such that $\Delta u'' - \pi'' + C(\pi'') \leq 0$. By (4) and the Inada assumption that $H'(0) = 0$ and $lim_{\tau \to 1} H'(\tau) = \infty$, it follows that $\tau_a = 0$ for all q_t. As $\Delta u > 0$ by Assumption 1, it follows from (5) that $\tau_b > 0$ for all q_t. It follows from (16) that if $\tau_a \neq \tau_b$ for all q_t, no interior SSE exists.

(iii) Imposing a coercion level π' corresponding to an interior SSE, $q^*(\pi') \in (0, 1)$, from an initial interior SSE, $q^*(\pi_0)$, will make q converge to $q^*(\pi')$.

Assume the population is in some interior $q^*(\pi_0)$, and at time $t = 0$ a $\pi' \neq \pi_0 : q^*(\pi') \in (0, 1)$ is imposed. As $\pi' \neq \pi_0$ and there is a unique interior SSE by part (ii) of the proof, the FOC conditions for an SSE cannot be fulfilled, i.e. $q_0(\Delta u - \pi' + C(\pi')) \neq (1 - q_0)(\Delta u + \pi')$ at time $t = 0$. This implies $H'(\tau_t^s) \neq H'(\tau_t^{ns})$; because of the Inada conditions on $H(\cdot)$ it follows that $\tau_t^s \neq \tau_t^{ns}$, and by (16) it follows that $q_1 \neq q^*(\pi_0)$. We define the following sequence of $q_0, q_1, \ldots, q_N$ values under π' as $\{q_0, q_1, \ldots, q_N, \pi'\} \equiv \{q_t\}_{\pi'}$.

We first establish that (*i*), any q_t in $\{q_t\}_{\pi'}$ will move in the direction of $q^*(\pi')$; (*ii*), no $q_t \in \{q_t\}_{\pi'}$ is equal to the absorbing state exterior SSE $q_t \in \{0, 1\}$; and finally (*iii*), that $q_t \to q^*(\pi')$.

(i) Any $q_t \in \{q_t\}_{\pi'}$ will move in the direction of $q^*(\pi')$.

By q_t moving in the direction of $q^*(\pi')$, we mean that if $q_t > q^*(\pi')$ then $q_t > q_{t+1}$ and if $q_t < q^*(\pi')$, then $q_t < q_{t+1}$.

First, note that as $q_t = q^*(\pi')$, as established in the first part (i) and (ii), it holds that $(1 - q^*(\pi'))(\Delta u - \pi' + C(\pi')) = q^*(\pi')(\Delta u + \pi')$. Suppose $q_t > q^*(\pi')$. It then follows that $(1 - q_t)(\Delta u - \pi' + C(\pi')) < q_t(\Delta u + \pi')$ which, by (4) and (5), implies $H'(\tau_a) < H'(\tau_b)$. It follows from the Inada condition of $H'(\cdot) > 0$ that this implies $\tau_t^s > \tau_t^{ns}$. Suppose $q_t < q^*(\pi')$, then the opposite holds. By (16) it holds that $q_t < q_{t+1}$ when $\tau_t^s < \tau_t^{ns}$ and $q_t > q_{t+1}$, if $\tau_t^s > \tau_t^{ns}$.

(ii) No $q_t \in \{q_t\}_{\pi'}$ is equal to the absorbing state exterior SSE: $q_t \in \{0, 1\}$.

We first show that an interior $q^*(\pi')$ implies positive levels of socialization for both groups at all $q_t \in \{q_t\}_{\pi'}$, then, we demonstrate that this implies no exterior $q_t \in \{0, 1\}$ is in $\{q_t\}_{\pi'}$.

[31] This argument holds mutatis mutandis for any $q^*(\pi'') = \frac{\Delta u''-\pi''+C(\pi'')}{2\Delta u''+C(\pi'')} = \frac{m}{n} \in (0, 1)$ and $q_t = \frac{m}{n}$. Hence, for any $\{\Delta u'', \pi''\}$ such that $q^*(\pi'') = \frac{\Delta u''-\pi''+C(\pi'')}{2\Delta u''+C(\pi'')} \in (0, 1)$, an internal SSE exists.

As $q^*(\pi') = \frac{\Delta u' - \pi' + C(\pi')}{2\Delta u' + C(\pi')} \in (0,1)$ it must hold that $\Delta u' + \pi' > 0$ and $\Delta u' - \pi' + C(\pi') > 0$. The FOC conditions, (4) and (5) and the Inada condition $H'(0) = 0$, imply that $\tau_a > 0, \tau_b > 0$ for any $q_t > 0$ when $\Delta u' + \pi' > 0$ and $\Delta u' - \pi' + C(\pi') > 0$. Hence, there will always be $\tau_t^{ns} > 0, \tau_t^s > 0$ under π' for all $q_{t-1} > 0$. As $q_0 \in (0,1)$, it follows that $\tau_a > 0$, $\tau_b > 0$ and $q_{t-1} > 0$ for all $q_t \in \{q_t\}_{\pi'}$.

From (16), $q_{t+1} = q_t + q_t(1 - q_t)(\tau^{ns} - \tau^s)$. We see that an exterior $q^* = 0$ or $q^* = 1$ cannot be reached from any interior $q_t \in (0,1)$ if $\tau^{ns} > 0, \tau^s > 0$.

(iii) $q_t \to q^*(\pi')$.

This proof applies Lemma 1, Lemma 2 and the definition of cultural substitution in Bisin and Verdier (bisin2001economics pp 303–307). Following the proof of Lemma 2 in Bisin and Verdier (2001), we show that socialization level τ, as a function of q_t, satisfies the definition of cultural substitution in Bisin and Verdier (2001). It then follows from Lemma 1 in Bisin and Verdier (2001) that this implies $q_t \to q^*(\pi')$.

We define $q_t^a \equiv q_t$ and $q_t^b \equiv 1 - q_t$, and denote a portion of a identity m, q^m. The requirements for cultural substitution on page 303 in Bisin and Verdier (2001) can be stated as: *(i)* $\tau^m = d^m(q_t^m)$, where $d^m(q_t^m)$ is a continuous function; *(ii)* $d^m(1) = 0$; and *(iii)* $d^m(q_t^m)$ is strictly decreasing in q_t^m.

(*i*) $\tau^m = d^m(q_t^m)$ is a continuous function.

From (4) and (5) it follows that:

$$\tau_a = H'^{-1}((1 - q_t)(\Delta u - \pi' + C(\pi')). \tag{20}$$

$$\tau_b = H'^{-1}(q_t(\Delta u + \pi')). \tag{21}$$

We first show that $H'^{-1}(\cdot)$ is defined. First, note that the Inada conditions $H'(\tau) \geq 0$, $H''(\tau) > 0$ and $H'(0) = 0$, imply that $H'(\cdot) > 0$ for all τ other than $\tau = 0$. The Inada condition $lim_{\tau \to 1} H(\tau) = \infty$ implies that $H'(\tau)$ maps from $[0,1) \to [0,\infty)$, $H'(0) = 0$ and $H''(\cdot) > 0$, implies $H'(\cdot)$ has a continuous positive derivative. Hence, for every q_t^m, $H'(\cdot)$ assigns a unique value; i.e. $H'(\cdot)$ is a one-to-one defined continuous inverse function $H'^{-1}(\tau^m)$, mapping from $[0,1) \to [0,\infty)$.

As everything inside $H'^{-1}(\cdot)$ in (20), (21) but q_t remains fixed for all $q_t \in \{q_t\}_{\pi'}$, and because $q^*(\pi') > 0$, implies that $\Delta u - \pi' + C(\pi') = K^{ns} > 0$, $\Delta u + \pi' = K^s > 0$; we can define $H'^{-1}(q^m K^m) \equiv d^m(q^m)$. As $q_t \in (0,1)$, we can write $\tau^{ns} = d^a(q_t)$ and $\tau^s = d^b(1 - q_t)$.

(*ii*) $d^m(1) = 0$.

Following from (2) and (3), parents are indifferent between choosing some infinitesimal amount of socialization and no socialization for $q_t \in \{0, 1\}$. By assumption in footnote 9 on page 18, we have assumed that it holds that $\tau^{ns} = 0$ for $q_t = 1$ and $\tau^s = 0$ for $(1 - q_t) = 1$, hence for $q^m = 1$ it holds that $\tau^m = 0$, i.e. $d^m(1) = 0$.

(*iii*) $d^m(q^m)$ is strictly decreasing in q^m.

We see from (20) and (21) that $\frac{\partial H'^{-1}((1-q_t)K_1)}{\partial q_t} < 0$ and $\frac{\partial H'^{-1}(q_t K_2)}{\partial (1-q_t)} < 0$ for all q_t. From the Inada assumptions $H''(\cdot) > 0$ and $H'(\cdot) \geq 0$, and that we have established that $H'^{-1}(\cdot)$ is continuously defined, it follows that $d'^m(q^m) < 0$ for all $q^m \in (0, 1)$.

The rest of the proof follows directly from Bisin and Verdier (2001) and Lemma 1. Inserting $\tau^{ns} = d^i(q_t)$ and $\tau^s = d^i(1 - q_t)$ into (16) and taking the continuous time limit and denoting the continuous rate of change $\dot{q}$, we attain equation (3) in Bisin and Verdier (2001) on page 303:[32]

$$\dot{q} = q(1 - q)[d^a(q) - d^b(1 - q)]. \tag{22}$$

From part *(i)* of the proof we have $\tau^{ns} > \tau^s$ when $q_t < q^*(\pi)$ and vice versa, hence it follows that $d^a(1 - q) - d^b(q) > 0$ for $q_t < q^*(\pi)$, and $d^a(1 - q) - d^b(q) < 0$ for $q > q^*(\pi)$. Similarly, from Lemma 3 we have $d^a(1 - q^*(\pi)) = d^b(q^*(\pi))$. Note that $\frac{\partial \dot{q}_t}{\partial q}\Big|_{q=0} = d^a(0) - d^b(1) > 0$ and $\frac{\partial \dot{q}}{\partial q}\Big|_{q=1} = d^b(0) - d^a(1) > 0$ because $d^m(\cdot)$ satisfies cultural substitution. As $q^*(\pi)$ is unique, and (22) continuously maps from q into $\dot{q}$, the basin of attraction of $q^*(\pi')$ under π' is (0, 1), which implies that $q_t \to q^*(\pi')$.

□

Proof of Lemma 2

Lemma 2: $q^*(\pi)$ is characterized by the following properties:

(I) a unique global or local maximum($\pi_{\overline{q}}$) and a unique global minimum($\pi_{\underline{q}}$)
or

(II) a unique global or local maximum($\pi_{\overline{q}}$), a local minimum ($\pi_{\underline{q}}$), and a global, potentially unique, minimum ($\pi' \in [\pi^e_{\underline{q}}, \pi_{\max}]$)
or

(III) a global minimum ($\pi'' \in (0, \hat{\pi})$, where $q^*(\pi'') = 0$).

In addition, there will always be a local or unique global maximum at $q^*(0) = \frac{1}{2}$.

Proof We prove the lemma by (*i*) demonstrating the existence of extremal points when $q^*(\pi) > 0$ for all $\pi \in [0, \pi_{\max}]$. We then show that the classes are exhaustive of all scenarios, by first (*ii*) noting what the sign of the derivative of $q^{*'}(\pi)$ over $[0, \pi_{\max}]$

[32] To see why the result is also valid for the discrete time case, see the discussion in Bisin and Verdier (2001) in footnote 9 on page 303.

must be; we then use this to show (*iii*) that any possible $q^*(0) + \int_0^{\pi_{\max}} q^{*'}(\pi)d\pi$ will place the functional form within either class I), II) or III). Finally, we show (iv) the uniqueness properties of the extremal points.

(i) Existence of extremal points

Suppose $q^*(\pi) > 0$ for all $\pi \in [0, \pi_{\max}]$. We show that this implies there exists a unique minimum in $(0, \hat{\pi})$, $\pi_{\underline{q}}$, and a unique maximum in $(\hat{\pi}, \pi_{\max})$, $\pi_{\overline{q}}$, where $\hat{\pi}$ is the turning point of $C''(\pi)$.

We start by showing there is a unique minimum in $[0, \hat{\pi})$. First, we show there exists at least one π such that $q^{*'}(\pi) = 0$ in $[0, \hat{\pi})$. It is established in the proof of Lemma 5 of the working paper version of this paper, Schøyen (2017), that:

$$q^{*'}(\pi) = \frac{(C'(\pi) - 2)\Delta u + \pi C'(\pi) - C(\pi)}{(C(\pi) + 2\Delta u)^2}. \tag{23}$$

It follows from that $q^{*'}(0) < 0$, which by (23) implies $(C'(0) - 2)\Delta u < C(0)$. Similarly, it follows that $q^{*'}(\hat{\pi}) > 0$ which by (23) implies $(C'(\hat{\pi}) - 2)\Delta u + \hat{\pi}C'(\hat{\pi}) > C(\hat{\pi})$. All functions are continuously defined by the C^2 assumption of $C(\pi)$ in (6), hence $(C'(\pi) - 2)\Delta u + \pi C'(\pi)$ and $C(\pi)$ must cross at $(0, \hat{\pi})$, giving $q^{*'}(\pi) = 0$ for some $\pi \in [0, \hat{\pi})$. We denote this π value $\pi_{\underline{q}}$. We now show that $\pi_{\underline{q}}$ is a unique value. Note that, following from Assumption 1, (6) and (9), we have $\Delta u > 0$, $C''(\pi) > 0$ for $\pi \in [0, \hat{\pi})$, and $C'(\pi) > 0$. This implies that the derivative of $(C'(\pi) - 2)\Delta u + \pi C'(\pi)$, $C''(\pi)\Delta u + C'(\pi) + \pi C''(\pi)$ is strictly larger than the derivative of $C(\pi)$, $C'(\pi)$, for all $\pi \in [0, \hat{\pi})$. This implies $C(\pi)$ and $(C'(\pi) - 2)\Delta u + \pi C'(\pi)$ can only cross once at $[0, \hat{\pi})$ and consequently, $\pi_{\underline{q}}$ is unique. Finally, we show that the unique $q^{*'}(\pi_{\underline{q}}) = 0$ in $[0, \hat{\pi})$ is a minimum. Note that the derivative of $C(\pi)$ is always smaller than the derivative of $(C'(\pi) - 2)\Delta u + \pi C'(\pi)$. Considering (23), we see that $q^{*'}(0) > 0$, $q^{*'}(\hat{\pi}) < 0$ and $0 < \hat{\pi}$ imply that $(C'(\pi) - 2)\Delta u + \pi C'(\pi)$ starts from an initial lower value at $\pi = 0$, surpasses $C(\pi)$ at $\pi_{\underline{q}}$, and is strictly larger than $C(\pi)$ for $\pi \in (\pi_{\underline{q}}, \hat{\pi}]$. Considering (23), we see this implies $q^{*''}(\pi) > 0$ for $\pi \in (\pi_{\underline{q}}, \hat{\pi}]$, $q^{*'}(\pi) > 0$ for any $\pi > \pi_{\underline{q}}$, and $q^{*'}(\pi) < 0$ for any $\pi < \pi_{\underline{q}}$. Hence, $\pi_{\underline{q}}$ is a unique minimum in $[0, \hat{\pi})$.

We now show the existence of a unique maximum point in $(\hat{\pi}, \pi_{\max}]$. Note that it follows from (6) and (9) that $C''(\pi) < 0$ for $\pi \in (\hat{\pi}, \pi_{\max}]$, $C(0) \geq 0$ and $C'(\pi) > 0$. This implies that $q'^*(\pi) = \frac{C''(\pi)(\Delta u + \pi) - C'(\pi)2(C(\pi) + 2\Delta u)}{(C(\pi) + 2\Delta u)^4} < 0$ for $\pi \in (\hat{\pi}, \pi_{\max}]$. From the proof of Lemma 5 of Schøyen (2017) it follows that $q^{*'}(\hat{\pi}) > 0$ and $q^{*'}(\pi_{\max}) < 0$. Hence, $q^{*'}(\pi)$ is continuous and strictly decreasing in $\pi \in (\hat{\pi}, \pi_{\max}]$ from strictly positive to strictly negative, hence there must be one, and only one, $\pi' \in (\hat{\pi}, \pi_{\max})$ such that $q^{*'}(\pi') = 0$. This π' is defined as $\pi_{\overline{q}}$, the unique maximum in $(\hat{\pi}, \pi_{\max}]$.[33]

(ii) The sign of $q^{*'}(\pi)$ First, note that from part (*i*) of the proof we have $\pi_{\underline{q}} < \hat{\pi} < \pi_{\overline{q}}$. From Lemma 5 of Schøyen (2017) and part (*i*) of the proof, it follows

[33] As $\lim_{\pi_{\max} \to \infty} \frac{\Delta u - \pi_{\max} + C(\pi_{\max})}{2\Delta u + C(\pi_{\max})} < 1$ for all $\Delta u \in [0, \infty)$, the exterior $q^*(\pi) = 1$ can never be reached; it consequently holds that $q^*(\pi_{\overline{q}}) \in (0, 1)$.

that $q^{*'}(\pi)$ is strictly increasing from $q^{*'}(0) < 0$ to $q^{*'}(\pi_{\underline{q}}) = 0$ and onward to $q^{*'}(\hat{\pi}) > 2$, and strictly decreasing from $q^{*'}(\hat{\pi}) > 2$ to $q^{*'}(\pi_{\overline{q}}) = 0$ and onward to $q^{*}(\pi_{\max}) < 0$. Hence, if $q^{*}(\pi) > 0$ for all $\pi \in [0, \pi_{\max}]$, then we note the following.

$$q^{*'}(\pi) < 0 \text{ for all } \pi \in [0, \pi_{\underline{q}})$$
$$q^{*'}(\pi) > 0 \text{ for all } \pi \in (\pi_{\underline{q}}, \pi_{\overline{q}})$$
$$q^{*'}(\pi) < 0 \text{ for all } \pi \in (\pi_{\overline{q}}, \pi_{\max}]$$

(iii) $q^{*}(\pi)$ will be characterized by functional form class (I), (II) or (III)

We first show that if $\pi_{\underline{q}}$ is not defined, then the functional form is of class (III). We than show that if $\pi_{\underline{q}}$ is defined it implies that $q^{*}(\pi)$ is characterized by class (I) or class (II). We then establish when $q^{*}(\pi)$ is characterized by class (I) or class (II).

Note that from part (*i*) of the proof we have $q^{*'}(0) < 0$ and $q^{*''}(0) > 0$. If $q^{*}(\pi'') = 0$ for some π'' in the interval $[0, \hat{\pi})$, where $q^{*'}(\pi) < 0$ such that $q^{*}(0) + \int_0^{\pi''} q^{*'}(\pi) d\pi = 0$ then, because $q^{*'}(\pi_{\underline{q}}) = 0$ and $q^{*'}(\pi_{\underline{q}}) > 0$ by definition, $\pi_{\underline{q}}$ is not defined. Then $q^{*}(\pi)$ is at a global minimum at this π'' and the functional form is of class (III).

If there is no $\pi'' \in [0, \hat{\pi}]$, while $q^{*'}(\pi) < 0$ such that $q^{*}(\pi'') = 0$ exists, then $q^{*'}(\pi') = 0$ where $q^{*}(\pi') > 0$ exists, and this π' is $\pi_{\underline{q}}$. As $q^{*'}(\pi) > 0$ for $(\pi_{\underline{q}}, \pi_{\overline{q}})$, $q^{*}(\pi) > 0$ for all $\pi \in [0, \hat{\pi}]$, it then follows from part (*i*) of the proof that there exist $\pi_{\underline{q}} \in (0, \hat{\pi})$ and $\pi_{\overline{q}} \in (\hat{\pi}, \pi_{\max})$.

We note that once $q^{*}(\pi) = 0$, the SSE for any π is zero. Thus, $q^{*}(\pi)$ ceases to change with π once it reaches 0. Hence, we can impose $q^{*'}(\pi) = 0$ for any $q^{*}(\pi) = 0$ such that we can define integrals of $q^{*'}(\pi)$ for $\pi \in [0, \pi_{\max}]$, even if $q^{*}(\pi) = 0$ for some $\pi \in [0, \pi_{\max}]$. Hence, we can write the integral of $q^{*}(\pi)$ over $[0, \pi_{\max}]$ for any functional form of $q^{*}(\pi)$ where $\pi_{\underline{q}}$ and $\pi_{\overline{q}}$ are defined as follows.

$$q^{*}(0) + \int_0^{\pi_{\underline{q}}} q^{*'}(\pi) \mathrm{d}\pi + \int_{\pi_{\underline{q}}}^{\pi_{\overline{q}}} q^{*'}(\pi) \mathrm{d}\pi + \int_{\pi_{\overline{q}}}^{\pi_{\max}} q^{*'}(\pi) \mathrm{d}\pi \tag{24}$$

We know the sign of $q^{*'}(\pi)$ in each interval from part (*ii*) of the proof. As $\pi_{\underline{q}}$ is defined, it follows that $q^{*}(0) + \int_0^{\pi_{\underline{q}}} q^{*'}(\pi) \mathrm{d}\pi > 0$.

We note the definition of $\pi_{\underline{q}}^{e}$ is $\pi_{\underline{q}} < \pi_{\underline{q}}^{e}$ and $q^{*}(\pi_{\underline{q}}^{e}) \equiv q^{*}(\pi_{\underline{q}})$. If $\int_{\pi_{\underline{q}}}^{\pi_{\overline{q}}} q^{*'}(\pi) \mathrm{d}\pi + \int_{\pi_{\overline{q}}}^{\pi_{\max}} q^{*'}(\pi) \mathrm{d}\pi \leq 0$, then, because all functions are continuous, $q^{*}(\pi_{\underline{q}}^{e})$ must be defined. Considering (24) and part (*ii*) of the proof, we see this implies that $q^{*}(\pi)$ has two minima, $\pi_{\underline{q}}$ and some $\pi' : \pi' \geq \pi_{\underline{q}}^{e}$, and one interior maximum, $\pi_{\overline{q}}$. This implies that the functional form falls within class II).

If $\int_{\pi_{\underline{q}}}^{\pi_{\overline{q}}} q^{*'}(\pi) \mathrm{d}\pi + \int_{\pi_{\overline{q}}}^{\pi_{\max}} q^{*'}(\pi) \mathrm{d}\pi > 0$, then there will be no $q^{*}(\pi')$ where $\pi' > \pi_{\underline{q}}$; i.e. $\pi_{\underline{q}}^{e}$ is not defined. Considering (24), we see this implies that all $\pi' > \pi_{\underline{q}}$ have the property $q^{*}(\pi') > q^{*}(\pi_{\underline{q}})$, i.e. $q^{*}(\pi)$ has only one minimum, $\pi_{\underline{q}}$, and one interior maximum, $\pi_{\overline{q}}$. This implies that the functional form lies within class I).

(iv) Properties of the extremal points

Following from the lemma and the sign of $q^{*'}(\pi)$ noted in part (*ii*) of the proof, there are five possible extremal points, two maximum points $\pi \in \{0, \pi_{\overline{q}}\}$, and three possible minimum points $\pi \in \{\pi', \pi_{\underline{q}}, \pi''\}$, where $\pi' \in [\pi^e_{\underline{q}}, \pi_{\max}]$ and $\pi'' \in (0, \pi_{\underline{q}})$. We here establish the properties of the points of importance in the lemma: $\pi \in \{\pi', \pi_{\underline{q}}, \pi_{\overline{q}}, \pi''\}$.

We first show the properties of $\pi'' \in (0, \pi_{\underline{q}})$. If follows from part (*iii*) of the proof that if π'' is defined, it implies $q^*(\pi'') = 0$, hence π'' is always a global minimum.

We now show when $q^*(\pi_{\overline{q}})$ is a local or global maximum. We have already established in part (*i*) that $\pi_{\overline{q}}$ is the only interior maximum point. From the sign of $q^{*'}(\pi)$ over $[0, \pi_{\max}]$ noted in part (*ii*) of the proof, it follows that the other possible maximum point lies at $q^*(0)$. If $q^*(0) < q^*(\pi_{\overline{q}})$, $\pi_{\overline{q}}$ is a unique global maximum; if $q^*(0) \geq q^*(\pi_{\overline{q}})$, then $q^*(\pi_{\overline{q}})$ is a local maximum.

We now show when $q^*(\pi_{\underline{q}})$ is a unique global, non-unique global or local minimum. Suppose that $\int_{\pi_{\underline{q}}}^{\pi_{\overline{q}}} q^{*'}(\pi)d\pi + \int_{\pi_{\overline{q}}}^{\pi_{\max}} q^{*'}(\pi)d\pi > 0$. From part (*iii*) of the proof this implies a functional form of class I), and $\pi_{\underline{q}}$ is the only minimum and hence a unique global minimum point. Suppose $\int_{\pi_{\underline{q}}}^{\pi_{\overline{q}}} q^{*'}(\pi)d\pi + \int_{\pi_{\overline{q}}}^{\pi_{\max}} q^{*'}(\pi)d\pi = 0$, then $\pi_{\underline{q}}$ is a non-unique global minimum, because it must then hold that $q^*(\pi_{\underline{q}}) = q^*(\pi_{\max})$. Suppose $\int_{\pi_{\underline{q}}}^{\pi_{\overline{q}}} q^{*'}(\pi)d\pi + \int_{\pi_{\overline{q}}}^{\pi_{\max}} q^{*'}(\pi)d\pi < 0$, then $\pi_{\underline{q}}$ is a non-unique local minimum because this implies there exists a π' such that $q^*(\pi_{\underline{q}}) < q^*(\pi')$.

We now show $q^*(\pi')$ where $\pi' \in [\pi^e_{\underline{q}}, \pi_{\max}]$ is a global minimum. As $q^{*'}(\pi) < 0$ for $\pi \in (\pi_{\underline{q}}, \pi_{\max})$ as established in (*ii*) of the proof, this minimum is unique global if $\pi' = \pi_{\max}$ and $q^*(\pi') < q^*(\pi_{\underline{q}})$. The minimum π' is non-unique global if $\pi' < \pi_{\max}$; this implies $q^*(\pi'''') = 0$ for all $\pi'''' \leq \pi'$. The minimum π' is also non-unique global if $q^*(\pi') = q^*(\pi_{\underline{q}})$ and $\pi' = \pi_{\max}$, as follows from the preceding discussion of the properties of $\pi_{\underline{q}}$. □

Proof of Lemma 3

Lemma 3: If a constraint affects coercion use under an exogenous constraint, $\rho \leq \pi^{NC}$ and $\rho \neq \pi^e_{\underline{q}}$, the following holds.

(*i*) $\pi^{EC} = \rho$ if and only if $\rho \notin (\pi_{\underline{q}}, \pi^e_{\underline{q}})$.

(*ii*) $\pi^{EC} = \pi_{\underline{q}} < \rho$ if and only if $\rho \in (\pi_{\underline{q}}, \pi^e_{\underline{q}})$.

Proof We first note the three possible scenarios of π^{EC} of $\rho \in [0, \pi_{\max}]$, and then demonstrate parts (*i*) and (*ii*) of the lemma.

If the authority is minimizing $q^*(\pi)$ by (1) and Lemma 2, it follows that if $\rho \neq \pi^e_{\underline{q}}$ and $\rho \leq \pi^{NC}$ then there are three different scenarios of $\rho \in [0, \pi_{\max}]$ as follows.

(I) $\rho \in [0, \pi_{\underline{q}}] \rightarrow \pi^{EC} = \rho$.

From part (*ii*) of the proof of Lemma 2, it holds that $q^{*'}(\pi) < 0$ for all $\pi \in [0, \pi_{\underline{q}})$. Hence, the minimal $q^*(\pi)$ for $\rho \in (0, \pi_{\underline{q}})$ is always equal to ρ.

(II) $\rho \in (\pi_{\underline{q}}, \pi^e_{\underline{q}}) \rightarrow \pi^{EC} = \pi_{\underline{q}} < \rho$.

By the proof of Lemma 2 part (*iii*), $\pi_{\underline{q}}$ is the minimum value in $[0, \pi_{\max}]$, unless $\pi^e_{\underline{q}}$ is defined. By definition $\pi^e_{\underline{q}}$ is a unique π value larger than $\pi_{\underline{q}}$, such that $q^*(\pi^e_{\underline{q}}) = q^*(\pi_{\underline{q}})$, which follows from (24) and the proof of Lemma 2 part (*ii*). Hence, for every $\pi' \in (\pi_q, \pi^e_q)$ it holds that $q^*(\pi') > q^*(\pi_{\underline{q}})$ and $\pi_{\underline{q}}$ must be the minimum of the open interval of $[0, \pi^e_{\underline{q}})$.

(III) $\rho \in (\pi^e_{\underline{q}}, \pi_{\max}] \rightarrow \pi^{EC} = \rho$.

From part (*ii*) of the proof of Lemma 2, $q^{*'}(\pi) < 0$ for all $\pi \in (\pi_{\overline{q}}, \pi_{\max}]$ because $\pi^e_{\underline{q}} \in (\pi_{\overline{q}}, \pi_{\max})$ any $\pi' > \pi^e_{\underline{q}}$ implies $q^*(\pi^e_{\underline{q}}) > q^*(\pi')$. By Lemma 2 it follows that $\pi = \pi_{\underline{q}}$ is the minimum of $\pi \in [0, \pi^e_{\underline{q}})$. As $q^*(\pi_{\underline{q}}) \equiv q^*(\pi^e_{\underline{q}})$ by definition in (10), the minimum $q^*(\pi)$ when choosing a $\pi^{EC} \in [0, \rho]$, where $\rho \in (\pi^e_{\underline{q}}, \pi_{\max}]$ is ρ.

Note that $\rho \in [0, \pi_{\underline{q}}]$ or $\rho \in (\pi^e_{\underline{q}}, \pi_{\max}]$ implies $\rho \notin (\pi_{\underline{q}}, \pi^e_{\underline{q}})$. Thus I) and III) can be combined so that the different scenarios of $\rho \in [0, \pi_{\max}]$ can be stated in the following.

$$\rho \notin (\pi_{\underline{q}}, \pi^e_{\underline{q}}) \rightarrow \pi^{EC} = \rho \tag{25}$$

$$\rho \in (\pi_{\underline{q}}, \pi^e_{\underline{q}}) \rightarrow \pi^{EC} = \pi_{\underline{q}} < \rho \tag{26}$$

Note that the lemma states that $\rho \neq \pi^e_{\underline{q}}$, and I) implies that $\rho = \pi_{\underline{q}} \rightarrow \pi^{EC} = \pi_{\underline{q}} = \rho$. Thus, (26) and (25) cover all possible scenarios of π^{EC} for $\rho \in [0, \pi_{\max}]$, which implies the following.

$$\pi^{EC} = \rho \rightarrow \rho \notin (\pi_{\underline{q}}, \pi^e_{\underline{q}}) \tag{27}$$

$$\pi^{EC} = \pi_{\underline{q}} < \rho \rightarrow \rho \in (\pi_{\underline{q}}, \pi^e_{\underline{q}}) \tag{28}$$

Part (*i*) of the lemma follows from (25) and (27). Part (*ii*) of the lemma follows from (26) and (28). □

Theorem of Restraints on Coercion A legitimacy-maximizing authority will restrain its use of coercion towards a non-state identity as a response to a constraint if and only if the efficacy of coercion varies and the restraint is in an inefficient interval of coercion.

Proof We first note that by definition an authority exhibiting restraint as a response to the constraint imposes a coercion level π^{EC} that is:

1. A response to a constraint; a π^{EC} different than its optimal adjustment without constraints, $\pi^{EC} \neq \pi^{NC}$.
2. A restraint; a π^{EC} strictly lower than its highest implementable level, $\pi^{EC} < \rho \leq \pi_{\max}$.

Second we note that the definition of varying efficacy of coercion, that is efficacy going from positive to negative and back to positive, only occurs under class II, this follows from Proposition 2.

Third, we note that there is only one inefficient interval of coercion for this model $(\pi_{\underline{q}}, \pi^e_{\underline{q}})$, this follows from Lemma 2 on functional form.

We first show that if part, $q^*(\pi)$ is of class II) and $\rho \in (\pi_{\underline{q}}, \pi^e_{\underline{q}})$ implies a restrain on coercion as a response to a constraint. We then show the only if part by first demonstrating that if $\rho \notin (\pi_{\underline{q}}, \pi^e_{\underline{q}})$ then there is no restrain on coercion. Finally, we show that if $q^*(\pi)$ of class I) or class III) and $\rho \in (\pi_{\underline{q}}, \pi^e_{\underline{q}})$ then π^{EC} is not a response to a constraint.

If $q^*(\pi)$ is in class II) and $\rho \in (\pi_{\underline{q}}, \pi^e_{\underline{q}})$, then from part (ii) of Lemma 3 $\pi^{EC} = \pi_{\underline{q}} < \rho$. Then π^{EC} is then a restraint as a response to a constraint since $\pi^{EC} = \pi_{\underline{q}} < \pi^e_{\underline{q}} < \pi^{NC}$.

If $\rho \notin (\pi_{\underline{q}}, \pi^e_{\underline{q}})$ and $\rho \neq \pi^e_{\underline{q}}$, then $\pi^{EC} = \rho$ by part i) of Lemma 3, hence π^{EC} is not a restrain. If $\pi^{NC} = \pi^{EC} = \{\pi^e_{\underline{q}}, \pi_{\underline{q}}\}$, then π^{EC} is not a response to a constraint.

If $q^*(\pi)$ is of class I) and $\rho \in (\pi_{\underline{q}}, \pi^e_{\underline{q}})$, then the optimal adjustment is equal with or without constraints, $\pi^{NC} = \pi^{EC} = \pi_{\underline{q}}$, thus π^{EC} is not a response to a constraint. If $q^*(\pi)$ is of class III) and $\rho \in (\pi_{\underline{q}}, \pi^e_{\underline{q}})$, then the optimal adjustment is equal with or without constraints $\pi^{NC} = \pi^{EC} = \pi'' < \rho$ where $q^*(\pi'') = 0$, consequently π^{EC} is not a response to a constraint. □

Proof of Lemma 4

Lemma 4: For any initial condition π_0, the DSE π^{DSE} is a coercion level equal to either:

(I) the unconfrontational level of coercion as an interior point of $\mathbf{I}_{\pi_0}$: $\pi^{\text{DSE}} = \pi_{\underline{q}}$
(II) a strategic constraint at the upper bound of $\mathbf{I}_{\pi_0}$: $\pi^{\text{DSE}} = \overline{\pi_{\text{fix}}}$

(III) the upper feasibility constraint at the bound of $\mathbf{I}_{\Pi}$: $\pi^{\text{DSE}} = \pi_{max}$.

Proof We define the set of SSE levels corresponding to the set of implementable coercion levels is denoted as $\mathbf{Q}_{\pi_0} \equiv \{q^*(\pi) : \pi \in \mathbf{S}_{\pi_0}\}$. $\mathbf{S}_{\pi_0}$ and $\mathbf{Q}_{\pi_0}$ will be non-empty for any π_0. We show that the π' corresponding to any minimum point of any $\mathbf{Q}_{\pi_0}$, which by definition is equal to π^{DSE}, will fall under either case (I), (II) or (III), hence the lemma.[34] First note that trivially, any minimum point in $\mathbf{Q}_{\pi_0}$ must correspond to a π^{DSE} in the interior of a subset of an $\mathbf{I}_{\pi_0}$, $\mathbf{i}$, or at the boundary of an $\mathbf{i}$.

Suppose the minimum of $\mathbf{Q}_{\pi_0}$ corresponds to an interior point in an $\mathbf{i}$. As established in Lemma 2, $q^*(\pi)$ has at most one interior minimum point, $\pi_{\underline{q}}$, and because $\rho'(q^*(\pi)) < 0$ always holds, $q^{*'}(\pi) = 0$ must hold at a minimum of $\mathbf{Q}_{\pi_0}$, corresponding to an interior minimum of $\mathbf{i}$. Thus, π^{DSE} must be equal to $\pi_{\underline{q}}$ and π^{DSE} fall under case (I).

Suppose the minimum of $\mathbf{Q}_{\pi_0}$ corresponds to a π^{DSE} that is the limit of a subset $\mathbf{i}$, and this limit is different from $\pi_{\max}$. π^{DSE} must be at an upper limit of $\mathbf{i}$, because at lower thresholds of $\mathbf{i}$ lowering π increases $q^*(\pi)$, as follows from proof of lemma 3. As the limit π^{DSE} is an upper limit different from $\pi_{\max}$, it implies there exists a $\pi^{\text{DSE}} < \pi' < \pi_{\max}$ such that π' is infinitesimally larger than the upper limit of the subset. As $\pi' \notin \mathbf{i}$, $\rho(q^*(\pi')) > \pi'$. As $\rho(\cdot)$ is assumed to be a continuous mapping with a continuous derivative, it cannot discontinuously jump from π^{DSE}, which is either on or over the 45-degree fix-point-line, to a point π' under the line, without crossing the fix-point-line.[35] Hence, the minimum of $\mathbf{Q}_{\pi_0}$ must correspond to an upper limit on the fix-point-line $\pi^{\text{DSE}} = \overline{\pi_{\text{fix}}}$, which falls under case (II).

Suppose the minimum of $\mathbf{Q}_{\pi_0}$ corresponds to an upper limit of a subset $\mathbf{i}$, then this limit is $\pi^{\text{DSE}} = \pi_{\max}$ and corresponds to case (III). □

Proof of Theorem of path dependency

Theorem of Path Dependency DSE is inherently given by the history of the polity if and only if there exist coercion levels that are sustainable but unimplementable from some initial conditions.

Proof First, we note that the statement "DSE is inherently given by the history of the polity" is equivalent to claiming "different initial conditions $\pi_0 = \overline{\pi}_0$ and $\pi_0 = \underline{\pi}_0$ give different dynamically stable equilibria: $\overline{\pi}^{IC}_{\overline{\pi}_0} \neq \underline{\pi}^{IC}_{\underline{\pi}_0}$."

Second we note that the statement "there exist coercion levels that are sustainable but unimplementable from some initial conditions" is equivalent to "there exist

[34] If there are several infimum points, any will correspond to a DSE, as the authority will not have any incentive to change π.

[35] The fix-point-line for $\hat{\rho}(\pi)$ is illustrated in Fig. 4 on page 34.

initial conditions $\overline{\pi}_0 \neq \underline{\pi}_0$, such that the set of implementable coercion levels from $\overline{\pi}_0$ or $\underline{\pi}_0$ differ, $\bar{\mathbf{I}}_{\bar{\pi}_0} \triangle \underline{\mathbf{I}}_{\underline{\pi}_0} \neq \emptyset$".

Since the theorem holds if and only if we can restate the theorem as follows: "If and only if there exist initial conditions, $\overline{\pi}_0 \neq \underline{\pi}_0$, such that the set of implementable coercion levels from $\overline{\pi}_0$ or $\underline{\pi}_0$ differ, $\bar{\mathbf{I}}_{\bar{\pi}_0} \triangle \underline{\mathbf{I}}_{\underline{\pi}_0} \neq \emptyset$, will different initial conditions $\pi_0 = \overline{\pi}_0$ and $\pi_0 = \underline{\pi}_0$ give different dynamically stable equilibria; $\overline{\pi}^{IC}_{\bar{\pi}_0} \neq \underline{\pi}^{IC}_{\underline{\pi}_0}$.

We first show that if $\bar{\mathbf{I}}_{\bar{\pi}_0} \triangle \underline{\mathbf{I}}_{\underline{\pi}_0} \neq \emptyset$, then $\overline{\pi}^{IC}_{\bar{\pi}_0} \neq \underline{\pi}^{IC}_{\underline{\pi}_0}$. We then show if $\overline{\pi}^{IC}_{\bar{\pi}_0} \neq \underline{\pi}^{IC}_{\underline{\pi}_0}$ then $\bar{\mathbf{I}}_{\bar{\pi}_0} \triangle \underline{\mathbf{I}}_{\underline{\pi}_0} \neq \emptyset$.

Suppose $\overline{\pi}_0 > \underline{\pi}_0$ and $\bar{\mathbf{I}}_{\bar{\pi}_0} \triangle \underline{\mathbf{I}}_{\underline{\pi}_0} \neq \emptyset$. By definition of the set of implementable coercion levels, there must be at least one π' such that $\pi' \notin \underline{\mathbf{I}}_{\underline{\pi}_0}$, but $\pi' \in \bar{\mathbf{I}}_{\bar{\pi}_0}$ because if this was not the case, then the sets would be the same sets; i.e. this is implied by $\bar{\mathbf{I}}_{\bar{\pi}_0} \triangle \underline{\mathbf{I}}_{\underline{\pi}_0} \neq \emptyset$. This implies that $q^*(\pi') < inf\{\mathbf{Q}_{\underline{\pi}_0}\}$, as $\rho(q^*(\pi))$ is monotonically increasing in $q^*(\pi)$ and π' cannot be reached from $\underline{\pi}_0$.[36] Suppose that the difference between the sets consists of this single coercion level π'. This $\pi' > sup\{\mathbf{I}_{\underline{\pi}_0}\}$ must then be equal to $\overline{\pi}^{IC}$, because π' must corresponds to $inf\{\mathbf{Q}_{\bar{\pi}_0}\}$. This π' is different from $\underline{\pi}^{IC}$ because π' is not in $\mathbf{I}_{\underline{\pi}_0}$.

Suppose the dynamically stable equilibria are different and that $\overline{\pi}^{IC}_{\bar{\pi}_0} > \underline{\pi}^{IC}_{\underline{\pi}_0}$. By definition $\overline{\pi}^{IC}_{\bar{\pi}_0}$ corresponds to $inf\{\mathbf{Q}_{\bar{\pi}_0}\}$. As $\rho'(\cdot) < 0$, it must hold that $q^*(\overline{\pi}^{IC}_{\bar{\pi}_0}) < q^*(\underline{\pi}^{IC}_{\underline{\pi}_0})$ because $\underline{\pi}^{IC}_{\underline{\pi}_0}$ can be implemented from $\overline{\pi}^{IC}_{\bar{\pi}_0}$, but $\overline{\pi}^{IC}_{\bar{\pi}_0}$ gives a lower $q^*(\pi)$ than $\underline{\pi}^{IC}_{\underline{\pi}_0}$, by definition. Hence, there must be at least one π' such that $\pi' \notin \underline{\mathbf{I}}_{\underline{\pi}_0}$ but $\pi' \in \bar{\mathbf{I}}_{\bar{\pi}_0}$, namely $\overline{\pi}^{IC}_{\bar{\pi}_0}$. By definition of the set of implementable coercion levels, this implies $\bar{\mathbf{I}}_{\bar{\pi}_0} \triangle \underline{\mathbf{I}}_{\underline{\pi}_0} \neq \emptyset$. □

Appendix 3: The set of implementable coercion levels

First the intuition of the insurrection constraint is discussed in Sect. 2.1. Then the following discussion provides some conjunctures about the set of implementable coercion levels under insurrection constraints with other mappings between q_t and the threshold level of insurrection, in "Appendix 3.2", and iterative processes for $\mathbf{I}_{\pi_0}$ where π can be set at any t, in "Appendix 3.3".

Appendix 3.1: Interpretations of the insurrection constraint

One possible reason the insurrection constraint has $\rho'(q^*(\pi)) < 0$, is to assume that increasing the size of non-state identity always increases their capability for committing a successful insurrection. The lower threshold for committing an insurrection

[36] $\mathbf{Q}_{\underline{\pi}_0}$ is defined in the proof of Lemma 4.

then follows from a higher probability of a successful outcome of an insurrection. Capability of attaining a successful outcome in an insurrection will grow with $q^*(\pi)$ for a wide number of applications, hence the assumption of $\rho'(q^*(\pi)) < 0$.

In applications of the model where military capability determines capability to perform a successful insurrection, the functional form of $\rho(q^*(\pi))$ is determined by the current military technology's ability to transform the share of *ns* identity individuals, $q^*(\pi)$, into military capability. The derivative of the insurrection constraint function at a particular SSE level, $\rho'(q^*(\pi'))$, will be determined by the relative labour intensity of military power. Assuming the insurrection constraint to be independent of SSE, $\rho'(q^*(\pi)) = 0$ for all $q^*(\pi) \in (0, 1)$ implies a military technology solely dependent on capital. A constant derivative, $\rho'(q^*(\pi)) = K$, for all $q^*(\pi) \in (0, 1)$ implies a military technology where every individual in the population has equal ability to exert military force and there is no scarcity of capital.

Appendix 3.2: Sufficiency of constraints on $\rho(q^*(\pi))$

We here discuss what requirement must be put on the model to ensure the insurrection constraint is not breached, given other relations between q_t and the threshold level of insurrection. We then discuss how this changes the set of implementable coercion levels.

The model considers an insurrection constraint on $q^*(\pi)$ rather than q_t. For a solution considering an insurrection constraint on $q^*(\pi)$ to be sufficient to imply that the solution would also hold for an insurrection constraint dependent on q_t, further restrictions are needed. The restrictions must ensure that whenever a coercion level, π', satisfying an initial insurrection constraint $\rho(q^*(\pi_0)) \geq \pi'$ is imposed, then this must imply that $\rho(q_t) \geq \pi'$ holds for all q_t in the sequence of q_t values in the convergence sequence from $q^*(\pi_0)$ towards $q^*(\pi')$. Following the notation in Lemma 1 part (*iii*), we denote this sequence of q_t values as $\{q_t\}_{\pi'}$. Now we discuss when the following criterion is met:

$$\text{If } \rho(q^*(\pi_0)) \geq \pi' \text{ and } \rho(q^*(\pi')) \geq \pi' \text{ then } \rho(q_t) \geq \pi' \text{ for all } q_t \in \{q_t\}_{\pi'}. \quad (29)$$

As $\rho(q^*(\pi))$ is monotonically strictly increasing, $\rho(q_t) \geq \pi'$ for all q_t is ensured if no $q_t \in \{q_t\}_{\pi'}$ is larger than the start of the convergence process; i.e. it must hold that $q^*(\pi_0) \geq q_t$, for all $q_t \in \{q_t\}_{\pi'}$. This is equivalent to a requirement of no-overshooting of $q^*(\pi_0)$; i.e. $q^*(\pi_0) \geq q_t$ for all $q_t \in \{q_t\}_{\pi'}$. Assuming a change in π occurring at time $t = 0$, then from (16) and requiring $q_t < q_{t+1}$ for any convergence path from $q^*(\pi_0)$ to $q^*(\pi')$, we derive that the cost function in socialization efforts must be sufficiently bounded for changes within $q_t \in (0, 1)$ such that:

$$\Delta_t \tau^m \geq \Delta_t \tau^m \Delta_{t+1} \tau^m + \Delta_{t+1} \tau^m \rightarrow 1 \geq \Delta_{t+1} \tau^m \left[1 + \frac{1}{\Delta_t \tau^m}\right]. \quad (30)$$

If (30) holds for all possible combinations of moving from one $q_t \in (0, 1)$ to another $q_{t+1} \in (0, 1)$, then (29) is satisfied. Hence, the requirement of no-overshooting is fulfilled as long as $|H(\tau_t) - H(\tau_{t+1})|$ is sufficiently bounded for changes in $q \in (0, 1)$.

Assume the insurrection constraint is dependent on a moving average $\tilde{\rho}(\overline{q}_{N,t})$, where $\overline{q}_{N,t} \equiv \frac{\sum_{i=0}^{N} q_{t-i}}{N+1}$. Furthermore, assume that the convergence process from $q^*(\pi_0)$ to $q^*(\pi')$ in (29) occurs within T periods. We know that $q^*(\pi')$ and $q^*(\pi_0)$ are sustainable, and it follows from proof of Lemma 1 part (*iii*) that any average will converge towards $q^*(\pi')$; i.e. it holds that $\overline{q}_{N,T} \to q^*(\pi')$ as $N \to \infty$. Hence, in the model as specified, (29) holds for an infinite moving average, i.e. $N = \infty$, infinite inertia. More simply put, (29) holds if military capability remains at $q^*(\pi_0)$ throughout the convergence process.

The smaller the N, the stricter the requirement on the convergence processes, and for $N = 0$, i.e. no inertia, (30) must always hold. As discussed on page 20, we have established that the convergence process will occur with every other generation in the process being above or below the SSE $q^*(\pi')$; i.e. $q_t < q_{t+2} < q^*(\pi') < q_{t+3} < q_{t+1}$. Hence, for any inertia process that can be described by a lag of more than two periods, $N \leq 2$, the requirements of the cost function in socialization efforts will be strictly weaker than under (30). Furthermore, because a shorter convergence time implies that $\overline{q}_{N,T}$ is closer to $\overline{q}_{N,0}$, we see that the requirement will be weaker if the convergence process is shorter, i.e. if T is low. Trivially for convergence in one period, $T = 1$, (29) always holds.

The discussion above has considered whether a constraint on the SSE can be considered a not sufficiently strict criterion to analyze which $q^*(\pi)$ an authority can dynamically reach. In other words, for another insurrection constraint dependent on q_t, there might be $q^*(\pi) \in \mathbf{Q}_{\pi_0}$ that the authority might not reach.[37] Furthermore, we have argued that it appears that the set of implementable coercion levels for an insurrection constraint dependent on $\overline{q}_{N,t}$, $\tilde{\mathbf{I}}_{\pi_0,N,T}$ will converge towards $\mathbf{I}_{\pi_0}$, as the inertia of military capability converges to infinity, $N \to \infty$, and the number of generations it takes to convergence between steady states converges to one, $\frac{1}{T} \to 1$.

Appendix 3.3: The set of implementable coercion levels when the authority can reset π at every t

In the specified model, the set of implementable coercion levels is given by what the authority can reach by setting π' in $q^*(\pi_0)$ and then resetting π once $q^*(\pi')$ is reached. Assume, as in "Appendix 2", that an insurrection constraint is dependent on q_t rather than $q^*(\pi)$, and that π can be reset at any t in the convergence sequence $\{q_t\}_{\pi'}$, defined in Lemma 4. The authority would then, potentially, be able to reach $q^*(\pi) \notin \mathbf{Q}_{\pi_0}$. This can arise as there might be q_t values in the convergence sequence, $\{q_t\}_{\pi'}$, from which the authority might be able to implement some π'' not implementable in $\mathbf{I}_{\pi_0}$, and thus reach $q^*(\pi) \notin \mathbf{Q}_{\pi_0}$. Investigating what states would then be reachable would require further inquiry into the extremal values of the convergence sequence $\{q_t\}_{\pi'}$. The states that would be sustainable, $\mathbf{I}_{\Pi}$, would not change, and there could still be limits regarding what is reachable from some initial condition; an authority could still be strategically constrained at an upper bound attractor fix-point

[37] $\mathbf{Q}_{\pi_0}$ is the set of SSE values corresponding to $\mathbf{I}_{\pi_0}$ defined in the proof of Lemma 4.

$\overline{\pi_{\text{fix}}}$. In other words, $\mathbf{I}_{\pi_0}$ might be different for other iterative processes, but it appears that all established results would hold qualitatively.

References

Acemoglu D, Robinson JA (2019) The narrow corridor: states, societies, and the fate of liberty. Penguin Press, London

Acemoglu D, Wolitzky A (2014) Cycles of conflict: an economic model. Am Econ Rev 104(4):1350–1367

Akerlof GA, Kranton RE (2000) Economics and identity. Q J Econ 115(3):715–753

Alesina A, Reich B (2013) Nation building. Technical report, National Bureau of Economic Research

Alesina A, Spolaore E (2005) The size of nations. MIT Press, Cambridge

Allardt E (1979) Implications of the ethnic revival in modern, industrialized society: a comparative study of the linguistic minorities in Western Europe. Societas Scientiarum Fennica

Barfield TJ (2010) Afghanistan: a cultural and political history. Princeton University Press, Princeton

Bennigsen A, Lemercier-Quelquejay C (1967) Islam in the Soviet union. Pall Mall Press, London

Besley T, Persson T (2019) Democratic values and institutions. Am Econ Rev Insights 1(1):59–76

Bisin A, Patacchini E, Verdier T, Zenou Y (2011) Formation and persistence of oppositional identities. Eur Econ Rev 55(8):1046–1071

Bisin A, Patacchini E, Verdier T, Zenou Y (2016) Bend it like beckham: ethnic identity and integration. Eur Econ Rev 90:146–164

Bisin A, Verdier T (2000) A model of cultural transmission, voting and political ideology. Eur J Polit Econ 16(1):5–29

Bisin A, Verdier T (2001) The economics of cultural transmission and the dynamics of preferences. J Econ Theory 97(2):298–319

Bisin A, Verdier T (2010) The economics of cultural transmission and socialization. Technical report, National Bureau of Economic Research

Bisin A, Verdier T (2017) On the joint evolution of culture and institutions. Technical report, National Bureau of Economic Research

Bowles S, Gintis H (2011) A cooperative species: human reciprocity and its evolution. Princeton University Press, Princeton

Boyd R, Richerson PJ (1988) Culture and the evolutionary process. University of Chicago Press, Chicago

Cantoni D (2015) The economic effects of the protestant reformation: testing the weber hypothesis in the german lands. J Eur Econ Assoc 13(4):561–598

Carvalho J-P (2013) Veiling. Q J Econ 128(1):337–370

Carvalho J-P, Dippel C (2016) Elite identity and political accountability: a tale of ten islands. Technical report, National Bureau of Economic Research

Carvalho J-P, Koyama M (2013) Resisting education. Technical report, University Library of Munich, Germany

Cavalli-Sforza LL, Feldman MW (1981) Cultural transmission and evolution: a quantitative approach, vol 16. Princeton University Press, Princeton

Choi S-W, Piazza JA (2016) Ethnic groups, political exclusion and domestic terrorism. Defence Peace Econ 27(1):37–63

Clemens WC Jr (2009) Culture and symbols as tools of resistance. J Baltic Stud 40(2):169–177

Conquest R (1970) The nation killers: the soviet deportation of nationalities. Macmillan, London

Conquest R, Case D (1991) Stalin: breaker of nations. Weidenfeld and Nicolson, London

Dippel C, Greif A, Trefler D (2016) The rents from trade and coercive institutions: removing the sugar coating. Rotman School of Management working paper, 2864727

Edgar AL (2004) Tribal nation: the making of Soviet Turkmenistan. Princeton University Press, Princeton

Fouka V (2016) Backlash: the unintended effects of language prohibition in us schools after world war I. Stanford center for international development working paper 591

Fouka V (2017) How do immigrants respond to discrimination? The case of Germans in the us during world war I. Am Polit Sci Rev, 1–18

Froese P (2008) The plot to kill god: findings from the soviet experiment in secularization. University of California Press, Berkeley

Fukuyama F (2018) Identity: the demand for dignity and the politics of resentment. Farrar, Straus and Giroux, New York
Gellner E (2008) Nations and nationalism. Cornell University Press
Gintis H, Van Schaik C (2013) Zoon politikon: the evolutionary origins of human political systems. In: Richerson PJ, Christiansen MH (eds) Cultural evolution: society, technology, language, and religion, vol 12. MIT Press, pp 25–44
Golman R, Loewenstein G, Moene KO, Zarri L (2016) The preference for belief consonance. J Econ Perspect 30(3):165–88
Greif A (2008) The normative foundations of institutions and institutional change (unpublished manuscript)
Greif A, Rubin J (2014) Endogenous political legitimacy: The English reformation and the institutional foundations of limited government. Working paper, Stanford University
Greif A, Schøyen Ø (2020) A theory of moral authority: moral choices under moral network externality (unpublished working paper)
Greif A, Tadelis S (2010) A theory of moral persistence: crypto-morality and political legitimacy. J Comp Econ 38(3):229–244
Huddy L, Sears DO, Levy JS (2013) The Oxford handbook of political psychology. Oxford University Press, Oxford
Jaeggi AV, Burkart JM, Van Schaik CP (2010) On the psychology of cooperation in humans and other primates: combining the natural history and experimental evidence of prosociality. Philos Trans R Soc Lond B Biol Sci 365(1553):2723–2735
Johnson ND, Koyama M (2013) Legal centralization and the birth of the secular state. J Comp Econ 41(4):959–978
Johnson ND, Koyama M (2019) Persecution and toleration: the long road to religious freedom. Cambridge University Press, Cambridge
Kirby D (1977) The baltic states 1940–1950. In Communist Power in Europe 1944–1949. Springer, pp. 22–35
Kreps DM (1997) Intrinsic motivation and extrinsic incentives. Am Econ Rev 87(2):359–364
Kula M (2005) Communism as religion. Total Mov Polit Relig 6(3):371–381
Kuran T, Sandholm WH (2008) Cultural integration and its discontents. Rev Econ Stud 75(1):201–228
Laitin DD (1998) Identity in formation: the Russian-speaking populations in the near abroad, vol 22. Cambridge University Press, Cambridge
Lenin VI (1909) The attitude of the workers' party to religion. Collect Works 15:402–13
Levi M (1997) Consent, dissent, and patriotism. Cambridge University Press, Cambridge
Levi M (1999) Death and taxes: extractive equality and the development of democratic institutions. In: Democracy's value. Cambridge University Press, Cambridge, New York, pp 112–131
Marshall RH, Bird TE, Blane A (1971) Aspects of religion in the Soviet Union, 1917–1967. University of Chicago Press, Chicago
Nash J (1953) Two-person cooperative games. Econometrica 21:128–140
Nexon D (2009) The struggle for power in early modern Europe. Princeton University Press, Princeton
North DC, Thomas RP (1973) The rise of the western world: a new economic history. Cambridge University Press, Cambridge
Northrop D (2001) Subaltern dialogues: subversion and resistance in soviet Uzbek family law. Slavic Rev 60:115–139
Philpott D (2001) Revolutions in sovereignty: how ideas shaped modern international relations. Princeton University Press, Princeton
Rae H (2002) State identities and the homogenisation of peoples, vol 84. Cambridge University Press, Cambridge
Rokkan S (1999) State formation, nation-building, and mass politics in Europe: the theory of Stein Rokkan: based on his collected works. Clarendon Press, Oxford
Rubin J (2011) Institutions, the rise of commerce and the persistence of laws: interest restrictions in Islam and Christianity. Econ J 121(557):1310–1339
Rubin J (2014) Printing and protestants: an empirical test of the role of printing in the reformation. Rev Econ Stat 96(2):270–286
Rywkin M (1990) Moscow's Muslim challenge: Soviet Central Asia. ME Sharpe, Armonk
Saleh M, Tirole J (2019) Taxing identity: theory and evidence from early Islam (No. 13705). CEPR discussion papers
Sambanis N, Shayo M (2013) Social identification and ethnic conflict. Am Polit Sci Rev 107(02):294–325

Schøyen Ø (2017) What limits the powerful in imposing the morality of their authority? NHH Department of economics discussion paper (18)
Stalin J (1975) Marxism and the national question. New Book Centre, New York
Sutherland NM (1988) The crown, the huguenots, and the edict of nantes. In: The Huguenot connection: the edict of nantes, its revocation, and early french migration to South Carolina. Springer, pp. 28–48
Theibault J (1997) The demography of the thirty years war re-revisited: Günther franz and his critics. German Hist 15(1):1–21
Thompson JW (1908) Some economic factors in the revocation of the edict of nantes. Am Hist Rev 14(1):38–50
Tilly C (1992) Coercion, capital, and European states, AD 990–1992. Blackwell, Oxford
Tilly C, Ardant G (1975) The formation of national states in Western Europe, vol 8. Princeton University Press, Princeton
Turchin P (2016) Ultrasociety: how 10,000 years of war made humans the greatest cooperators on earth. Beresta Books, Chaplin, CT
Verdier T, Zenou Y (2018) Cultural leader and the dynamics of assimilation. J Econ Theory 175:374–414
Wheeler G (1967) The muslims of central asia. Probs Commun 16:72
Wilson PH (2009) The thirty years war: Europe's tragedy. Harvard University Press, Cambridge
Woodburn J (1982) Egalitarian societies. Man 17:431–451