Legitimacy under dual moralities

Øivind Schøyen
Economics Department Norwegian School of Economics(NHH)
Centre for Experimental Research on Fairness, Inequality and Rationality (FAIR)

❖

Contents

1. Acknowledgments

I am a very luck man; throughout my Ph.D. studies by being surrounded by many inspiring and kind people who have supported me and helped me make this academic journey possible.

First and foremost, I would like to thank my supervisors: Bertil Tungodden in the Economics Department of the Norwegian School of Economics (NHH) and the Centre for Experimental Research on Fairness, Inequality and Rationality (FAIR), and my co-supervisor Avner Greif in the Economics Department at Stanford University. I would like to thank Bertil Tungodden for his hard work, good advice and excellent supervision. Thank you very much, Bertil, for also teaching me the value of clean and concise experiment design, writing and sharp thinking. I am thankful to Avner Greif for recognizing my potential from my master's dissertation, for his engaging conversation, inspiring words and generosity. Thank you very much, Avner, for all your generous help. I am also a very big fan of the work and sharp thinking of both my supervisors. I appreciate their contributions to economics, which have been stimulating to read and greatly beneficial to my thinking.

Second, I thank Xianwen Chen at The Norwegian Institute for Nature Research, for excellent co-authorship on the third chapter.

Third, there are many people at NHH who I have had the pleasure of working with. Little did I know upon arriving at the school how much I would learn from and appreciate the people there. In particular, I had the pleasure of interacting with faculty at the school: Kjetil Bjorvatn, Alexander Wright Cappelen, Thomas de Haan, Armando José Garcia Pires, Eirik Gaard Kristiansen, Ola Honningdal Grytten, Katrine Vellesen Løken, Linda Nøstbakken, Trond Olsen, Kjell Gunnar Salvanes and Erik Øiolf Sørensen. There are also so many students whom I have

had the pleasure of interacting with, too many to name. A few of them are Agnes, Bendik, Ceren, Elias, Felix, Ingvild, Oddleif, Seidal, Simen and Thor Andreas. I would like to thank all the excellent staff and research coordinators at NHH; in particular, I would like to thank Anne Liv Scarce and Vivienne Bowery Knowles for their research assistance and the research coordinators at FAIR: Ranveig, Ida, Janina and Ingeborg. I would also like to thank Bjarte Grønner and Anne Jorge Finnigan at the section for research administration for help on applications to the Norwegian Research Council.

There are also some kind and generous people at the University of Bergen (UIB) that I would like to thank: Tom Grimstvedt Meling, Bjørn Sandvik, Eirik Andre Strømland, Rune Jansen Hagen at The Economics Department for comments and conversations, Jørgen Pedersen at the Seminar for Political Philosophy at the Philosophy Department, and David Lara Arango and Erling Moxnes for interesting conversations at the System Dynamics Group at the Department of Geography.

Students and faculty from various universities have also been helpful in conversations about my research; a few I would particularly like to thank are Thomas Barfield, Gary Charness, Erik Eikeland, Andreas Espegren Masvie, Scott Gates, Alex Imas, Siri Isaksson, Jo Thori Lind, Karl Ove Moene, Andy Michael Martinangeli, Torben Mideksa, Anna Pauls, Henrik Singstad and Daniel Spiro. I spent one of the best years of my life at the University of California, Berkeley. The people who contributed to this rewarding year deserve recognition, in particular, Zarek Brot-Goldberg, Shachar Kariv, Jonas Tungodden and Pawel Gniewek. I would like to thank NHH, the FAIR Centre and the Economics Department at UC, Berkeley for giving me this excellent opportunity. I would also like to thank friends and staff at the International House, Berkeley which forever will feel like home. I also had some excellent shorter visits at the Institute for the Study of Religion (IRES) at Chapman University in Orange

County, The Association for the Study of Religion, Economics, and Culture (ASREC) conference in Boston, The 68 degree North Conference at Svolvær and Stanford University; in particular, I would like to thank Avner Greif, Bertil Tungodden, Jared Rubin and Larry Iannaccone for making these trips possible.

I have had the pleasure of completing and auditing several excellent courses and I would like to thank my teachers: Joshua Angrist (Advanced Labor Econometric Methods at NHH), Miguel Costa-Gomes (Experimental Markets at UIB), Tore Ellingsen (Institutional Economics at NHH), Armin Falk (Experiments and Moral Behavior at NHH), Sjur Flåm (Advanced Game Theory at UIB), Matthias Sutter (Experimental Economics at NHH), Uri Gneezy (Behavioral Economics at NHH), Noah Forman (Discrete Mathematics at UC Berkeley), Larry Iannaccone (Economics of Religion at Chapman University), John List (Experimental Economics at NHH) and Marco Ottaviani (Informational Economics at University of Oslo). I would also like to thank the teachers of the Economics Ph.D. course package at NHH.

I also gave a number of presentations during my Ph.D., and I thank the people who attended my presentations, provided valuable feedback and the institutions that made these presentations possible: ASREC, Christian Michelsens Institute for Research for Development and Justice, FAIR Centre, IRES, Norwegian Association of Economists, Nordic Conference of Development Economics, Political Theory Seminar in the Philosophy Department and the System Dynamics Group in the Geography Department at UIB, Strategic Analyses and Joint Systems Division at the Norwegian Defence Research Establishment, Department of Strategy and Management at NHH, and the economics departments at NHH, Norwegian University of Life Sciences, Oslo Metropolitan University and University of Oslo.

I am very grateful for financial support from my family, the Norwegian Research Council through the "Working life and migration grant" (grant number 236995),

"Understanding paternalism grant" (grant number 262636) and Centre of Excellence grant given to the FAIR Centre (grant number 262675), "Skipsreder J.R. Olsen og hustru J.G. Olsens legat til NHH", FAIR Centre, NHH, Department of Economics at NHH and John Templeton Foundation. I would also like to thank Public Art Norway (KORO) for holding the "Wittgenstein on Vacation" seminar in Skjolden, Norway and sponsoring my inspiring trip there.

Finally, I would like to thank my friends and family; I have so much love and respect for you. My friends have been there for me throughout my Ph.D. and I would like to thank all of them; in particular, Eirik Paulsberg, Bjarne Gjelland and Karin Lillevold, who helped me in different important ways throughout my Ph.D. My family has supported me, loved me and helped me throughout my Ph.D.: Thor Øivind Jensen, Elisabeth Schøyen Jensen and May Schøyen.

Thank you all very much; I am very grateful.

2. Introduction

During the Soviet Union period (1922–1991), administrators in Moscow sought to modernize its predominantly Muslim Central Asian regions. 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 legitimize "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, Bird and Blane, 1971). Implicitly, the council signalled that they were sympathetic to Stalin's cause, but that he would not succeed without the assistance of Allah (Froese, 2008).

This example illustrates the subject of this dissertation, legitimacy under dual moralities: the moralities are Islam and Communism while the question of legitimacy is the Soviet rule in Central Asia. The concept of morality can be contrasted with our tendency for sympathy: sympathy is our innate tendency to sympathize with the needs of others, while our notions of morality are abstract ideas that guide us to balance the needs of several parties when these are in conflict (Tomasello, 2016). While sympathy is a tendency we are born with, the content of our moralities is learned through socialization. Different people from different backgrounds, such as Russians and Kyrgyz, have often internalized different moralities, such as Communism and Islam. Legitimacy of an outcome implies that it adheres sufficiently to a morality. In our opening example, the Muslim scholar points out that the enforcement of Soviet

rule assures the outcome of public order; a natural interpretation would be that he sees Soviet rule in Central Asia as sufficiently adhering to Islam as to be legitimate.[1] Where people with different moralities interact, rules need to account for different notions of what is legitimate. This is important for avoiding conflict and ensuring that formal rules are enforced. Enforcing rules restricting behaviour that is seen as legitimate can often be seen as a transgression by the restricted, which can often lead to conflict. Rules restricting behaviour that is considered to be legitimate tend not to be enforced. Thus, to create functioning formal rules without causing conflicts in populations with different moralities, we need to understand legitimacy under dual moralities.

This dissertation is about the dynamics that may arise when actors seek to socialize others into their moralities, interpret whether moralities, or selfish motives, motivate actions, or enforce their views of what is legitimate. All chapters of this dissertation consider interactions between people with two different and distinct moralities prescribing mutually excluding outcomes as legitimate.

In my opening example, the Soviet and Muslims authorities both saw it in their interests to co-operate (Conquest, 1970); hence, they had incentives to find common ground and did so to a large extent. It is, however, often the case that people with different moralities do not find such common ground, either because their leaders do not have incentives to encourage them to do so, or simply because it can often be a major challenge. A thematic common thread in the chapters of this dissertation is to explain why this is the case; the dissertation focuses on mechanisms that help explain why interaction among people with different moralities is often a major, and

[1] As the observant reader might have deduced, an alternative title of my dissertation, "Legitimacy under dual moralities" could be "Outcomes sufficiently adhering to two abstract ways of thinking about balancing sympathy for others".

enduring, cause of friction and conflict among them.

The first chapter explores the consequences of people disliking being pressured to change their own morality. The second chapter establishes that, and offers insight into why, people with different moralities often consider others' intentions as selfish. The third chapter shows that people are interested in enforcing solutions adhering to their own moralities, independently of what those affected by their decision find legitimate. A methodological common thread throughout the dissertation is that the chapters create concepts and analytic tools that structure our understanding of legitimacy in dual morality settings, before it reviews empirical material through the lens of developed theories.

The theories developed in this dissertation are made under the assumption that when people with different moralities interact certain dynamics arise with some regularity. This type of dynamics can be understood by abstractions that allow insight to be transferred from one example to another. An example of such regularity is our tendency to discretely categorize each other into groups. Although the content of moralities might be dynamic and vary from person to person, people's mutual perceptions and categorizations of each other's moral views are often stable and distinct Tomasello (2014). If their categorization into discrete separable categories, such as "Muslim" or "Communist", remains stable, complex, yet predictable, dynamics might arise. Thus, understanding and describing this complex dynamic for one case might inform understanding of another, i.e., the dynamics arising between "Liberal" and "Conservative" might also arise between "Protestant" and "Catholic". The models are a stylization of the mechanisms to explain the phenomena that occur in the real world. They are built as logical consequences that arise from explicit assumptions; the models are not meant to be a holistic picture of every relevant parameter determining a phenomenon. Thus, the conclusions of the model hold as

long as the assumptions hold and the insights are relevant as long as the mechanism is the relevant driving mechanism behind a phenomenon. I now describe the three chapters in more depth.

The first chapter, "What limits the powerful in imposing the morality of their authority?", asks why moralities in a society are not always a direct function of the wishes of its rulers. The chapter offers a theory of why and how this can hold, even in the very long run, by building on a surprisingly simple assumption: coercion resentment. Coercion resentment is the assumption that people have an intrinsically negative reaction to being coerced, making coercion a potentially counter-productive measure. The chapter explores the consequences of coercion resentment in a formal model of how rulers can use coercion to discourage socialization of moralities other than the one upon which the ruler has built his legitimacy. The model focuses on the dynamics arising from the interaction between three decisions: the ruler's choice regarding how much coercion to apply; and the decisions of the ruled regarding how much effort to put into socializing others into their morality and whether or not to try to remove the ruler through committing an insurrection. Two historical periods are presented through the lens of the model: the Counter-Reformation in early modern France and the Holy Roman Empire (1517–1685), and the Soviet secularization project (1922–1991).

The second chapter, "Suspicious minds and views of fairness", investigates how we understand the motivation of people with moralities different to our own. The chapter opens by structuring, in a simple model, our understanding of how we interpret intentions behind actions. Central to this chapter is the concept of suspicious attribution: our tendency to attribute behaviour not adhering to our own morality to selfishness. I show that suspicious attribution can arise from under-estimating how many have a morality differing from our own, projection bias; and overestimat-

ing the selfishness of people with moralities different to our own, out-group stereotypes. I then present results in support of suspicious attribution from an incentivized laboratory setting. Participants in my experiment display both suspicious attribution, projection bias and out-group stereotypes, and my data suggest these are what cause suspicious attribution. Finally, I discuss how this mechanism can be central to understanding polarization of attitudes, and how prosocial individuals legitimize engagement in group conflicts.

The third and final chapter, "Paternalist motivation: An experimental test", was written together with Xianwen Chen. The chapter examines whether we take the morality of those affected into account when deciding whether or not to impose states adhering to our own morality upon others. The chapter offers the terms "non-paternalist" or "paternalist motivation" to differentiate what motivates us to impose states upon others. Non-paternalist motivation is defined as a desire to impose states upon others only if those affected think it constitutes a legitimate outcome, while paternalist motivation is motivation to implement our own view regardless of the views of those affected. The chapter then offers an experimental test of whether the moralities of the affected are relevant when we decide whether or not to enforce our moralities. We conduct an experiment and find that our participants have paternalist motivation: they are willing to pay to implement outcomes adhering to their moralities, but they do not take the moralities of the affected into account. We discuss how this finding has implications for collective choice in groups where moralities vary.

References

Conquest, Robert. 1970. *The Nation Killers: The Soviet Deportation of Nationalities*. Macmillan London.

Froese, Paul. 2008. *The Plot to Kill God: Findings From the Soviet Experiment in Secularization*. University of California Press.

Marshall, Richard H, Thomas E Bird and Andrew Blane. 1971. *Aspects of religion in the Soviet Union, 1917-1967*. University of Chicago Press.

Tomasello, Michael. 2014. *A natural history of human thinking*. Harvard University Press.

Tomasello, Michael. 2016. *A natural history of human morality*. Harvard University Press.

Chapter 1

What limits the powerful in imposing the morality of their authority?

Abstract

This paper models a game between an authority seeking to implement its preferred morality, and a parental generation seeking to socialize a younger generation into their own morality. The authority chooses a coercion level for adhering to the non-state morality, whereupon the parental generation chooses whether to insurrect and, if not, how much to invest in socialization. The novel feature of this paper is that we formalize and explore the consequences of an intrinsic negative reaction to coercion: coercion resentment. The key result is to show the necessary micro level assumptions for an inefficient interval of coercion that can account for authorities choosing to restrain their use of coercion. Furthermore, the paper characterizes the socialization and insurrection preferences required for long-run equilibrium to be path dependent. Two historical periods are presented through the lens of the model: the Counter-Reformation in early modern France and the Holy Roman Empire (1517–1685), and the Soviet secularization project (1922–1991).

1. Introduction

Polities generally seek to have legitimacy; that is, to rule in alignment with the internalized moralities of its population.[1] A crucial dimension of state legitimacy, and the focus of this paper, is whether the values of the polity, on which its formal institutions and moral right to rule are built, are aligned with the moralities of its population.[2] One way of attaining legitimacy is by using extrinsic incentives, generally referred to as coercion. This approach may, however, invoke an intrinsic counteraction, making coercion potentially counter-productive. The paper embeds this micro assumption, referred to as coercion resentment, into an overlapping generations model of moralities where an authority seeks to maximize the prevalence of its preferred morality by using coercion. The model analyzes how opposing effects of the extrinsic incentives to comply, and the intrinsic incentives to resist, determine the prevalence of different moralities.

Greif and Tadelis (2010) poses the question "Why do the powerful often fail to promote the morality of their authority?". In other words, what are the mechanisms behind moral persistence in the face of hostile institutional environments? If people simply choose to internalize the morality that gave them the highest extrinsic utility, moralities, and subsequently group identities, would simply be a function of the institutional environment. This would imply that everyone holds the most opportune

[1] A morality can be understood as a vector of beliefs and values that is internalized and embedded in a person: examples are political ideologies, religious or ethnic identities.

[2] Another important dimension is whether the state works to fulfill the values on which it builds its institutions, or whether it serves the interests of individuals who control the state, commonly referred to as corruption (Nye, 1967).

morality; in most cases, the morality aligned with the ruling regime. The dynamics of moralities such as religious, national or ideological identities could then be ignored in political economy analysis and at most, be treated as a rigidity. Assuming moralities are a passive function of extrinsic incentives would, however, account poorly for the persistence of minority identities such as the Jews in Europe, states' investments in costly nation building and foreign nationals' voluntary participation in perilous group conflicts such as the Spanish Civil War (1936–1939).

These historical instances illustrate that intrinsic reactions play an important role in the dynamics of legitimacy and state development, both in the short and long term. Minority moralities in hostile institutional environments can be remarkably persistent, as demonstrated by the historical evidence presented in Greif and Tadelis (2010) of Jews in Medieval Spain, while other historical examples, such as David Laitin's study of the Russian diaspora in the former USSR (Laitin, 1998), show quick adoption of new beliefs, norm sets and national identities, pointing to a rapid change in internalized values.

This paper's main contribution is to build a micro-founded model decomposing the effect of coercion aimed at changing moralities, into extrinsic and intrinsic reactions. The extrinsic reaction to coercion is a reduction of group identification as a response to incentives, while the intrinsic reaction to coercion is a strengthening of in-group identification and out-group resentment within the group being targeted. The model assumes that authority attempts to force people to change their moralities will invoke a resentment towards the authority, making certain levels of coercion counter-productive to attaining legitimacy.

A premise of the model is that authorities seek to maximize legitimacy. 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 staying in power, reduces enforcment costs and expands the possibility frontier of imposing policy (Greif, 2008), and increases willingness for altruistic behaviour, such as conscription (Levi, 1997) or payment of taxes (Levi, 1999). The 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 morality 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. Furthermore, a population with homogenous moralities enables central policy-making (Tilly, 1992); indeed, services such as law and policing hinge on, and grow out of, common sets of norms and values.

In the short term, the most obvious way to gain legitimacy is to take norms and values as given, and rule in accordance with the prevailing majority morality. To authorities of polities with heterogenous moralities, this implies making compromises between moralities where they are incompatible, typically at the cost of reduced legitimacy (Johnson and Koyama, 2013). A state might, however, enhance its legitimacy by increasing the portion of the population with internalized norms similar to those of its institutions. This can be done either by application of "sticks": disincentives and coercion, or "carrots": increasing the incentives of belonging to the authority's

morality.[3] This article focuses on the "stick" approach, coercion, and how it invokes an intrinsic negative reaction, making it a potentially counter-productive measure.[4]

The model develops necessary assumptions for analyzing the equilibrium coercion level and morality prevalence in the overlapping generations model of Bisin and Verdier (2000, 2001). This equilibrium is given as a function of parental preferences for their child adhering to their morality, and the strength and functional form of coercion resentment. A key result of this analysis is to show that authorities will only restrain use of coercion when there exists an inefficient interval of coercion, which is shown to imply a non-linear response to coercion. The paper then analyzes the dynamic problem of what constitutes equilibrium coercion level when the authority can iteratively reset the coercion level and the prevalence of minority morality determines an insurrection constraint on coercion use. We explore the dynamic property of states in coercion reliance, defined as an inability of an authority to decrease coercion, as doing so will increase minority prevalence thus increasing their insurrection capability and triggering an insurrection. The key result of this analysis is to show when responses to coercion create path dependency i.e. the insurrection and socialization preferences needed for equilibrium minority prevalence to be dependent on the history of the polity.

[3] Other measures include increasing socialization and easing communication by creating common standards, i.e., through building of roads, standardizing the language, creating common school systems and investing in common symbols.

[4] To the extent that "carrots", i.e., positive incentives, invoke a negative reaction among members of the non-state morality, the analysis generalizes to authorities imposing positive incentives for adhering to their morality.

The paper presents anecdotal historical evidence to demonstrate macro level restraints on coercion use. First, we review the Counter-Reformation in early modern France (1517–1685) and the Holy Roman Empire (1517–1648). The early modern French kings and the Holy Roman emperors built their legitimacy on the Catholic faith. The spread of Protestantism following Luther (1517) posed a direct threat to their program of state consolidation. As a response to this introduction of religious heterogeneity, they embarked on programs of homogenization. We argue that in this period, only unconfrontational or strongly coercive policies were stable over time. This supports model predictions that authorities restrain coercion use, and that any long-term stable coercion level must not give the authority any incentive for gradual increases in coercion. The paper then presents a brief comparative study of European early modernity before and after the Peace of Westphalia (1648) through the lens of the model; it shows how the change of international institutions affects constraints on use of coercion and consequently, minority prevalence.

Furthermore, we review evidence from Soviet secularization policies (1922–1991) towards the Christians and Muslims in the USSR. The Soviet Union sought to increase its legitimacy by increasing support for communism and diminishing the importance of religion. This secularization project was conducted in a comparatively more cautious way in regions where cultural differences were larger, recognizing the potential counter-productiveness of secularization attempts, in line with the proposed micromechanism of coercion resentment.

The paper develops as follows: the remainder of Section 1 reviews the related literature, Section 2 presents the overlapping generations model of Bisin and Verdier (2000, 2001) and Section 3 expands the basic model to include a legitimacy-maximizing au-

thority, coercion resentment and an endogenously determined insurrection constraint. Section 4 shows how the macro predictions of the model fit the Soviet secularization project and the Counter-Reformation in early modern France and the Holy Roman Empire. Section 5 concludes the paper and raises questions that can be investigated in future extensions of the theory. The appendices contain proofs, and some further analysis and interpretations of the model.

1.1. Related literature

The model's critical micro assumption is coercion resentment: individuals are assumed to react negatively towards the authority in response to coercion. More specifically, it is assumed that at least some level of coercion to hold a morality will cause individuals to respond by increasing their investment in socializing this morality. Why individuals act in such a way can be understood from different strands of the literature. Three main perspectives are reciprocity, fulfilling internalized norms and increased investment in social motives to help the group face a common external threat.

Reciprocity: coercion resentment can be understood 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 who have internalized the coerced morality and feel that the authority has harmed their group, wish to punish the group associated with the coercion through activities aimed at stopping the authority's influence.

Salience of fulfilling internalized norms: coercion resentment might also be understood as increased salience of acting in accordance with internalized norms. The au-

thority, and indirectly, the individuals aligned with state morality, become a salient enemy of non-state morality if they appear as having hostile intentions. The need to act in line with non-authority internalized norms will involve confronting individuals of state morality and stopping the spread of their morality, once they are conceived as being a threat. 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 opposing his values, as this will help defend his internalized values.

Social motives: a threat from an external foe increases in-group identification. This finding has a long-standing tradition and has solid empirical support in the social psychology literature (Huddy, Sears and Levy, 2013). As coercion of non-state morality increases, the authority will be seen as a threat to the non-state morality group. This increased external threat invokes an emotional reaction that triggers investment in social identity activities for individuals who have internalized non-state morality. The presence of a threat to the group increases in-group identity and strengthens hostility towards the out-group. The out-group threat effect is documented to increase a number of different group-related behaviours, including increased investment in socialization (Huddy, Sears and Levy, 2013).[5] Finally, once coercion is imposed on a morality, defying the coercion and acting in accordance with the coerced morality become costly, and can hence be used as a credible social signal of intrinsic motivation.

[5]Although strengthening of group identity is theoretically different from a utility loss of children adhering to an opposing morality, the implications are equal for the purpose of this study: a society of two mutually excluding moralities.

The paper draws on classical political science analysis of the state's role in moral dynamics. This literature initially focused on cultural unification into nation states, arguing that the relatively high pre-existing (pre-990 A.D.) homogeneity of morality in Europe contributed to Europe's relatively rapid state consolidation (Tilly and Ardant, 1975), and later focused on the survival and persistence of minority cultures through mechanisms of cultural resistance (Allardt, 1979; Rokkan, 1999).

The paper relates to four strands of the economics literature: social economics, group conflict, state legitimacy and path dependency in societal outcomes. The model is an expansion of the social economics model by Bisin and Verdier (2000, 2001), where overlapping generations transfer moralities and the prevalence of each morality is determined by parental investment in socialization. Models in social economics have addressed the role of cultural persistence through differences in socialization investment by mechanisms such as oppositional culture (Bisin et al., 2011), bias in education systems (Carvalho and Koyama, 2013) and social signalling of identity (Carvalho, 2013). Social economics models have generally not focused on actions of state actors or individuals' relation to a state (Bisin and Verdier, 2010). Following Greif and Tadelis (2010), this paper extends the author's master's thesis (Schøyen, 2011) and is novel in making a connection between the policies of a state authority and the prevalence of non-state minority moralities. Greif and Tadelis (2010) introduce an authority that controls the institutional environment to maximize the morality on which it builds its legitimacy, into the Bisin and Verdier (2000, 2001) framework. This paper extends Greif and Tadelis (2010) by letting the agents in the model intrinsically react to coercion. In contrast, the agents in the model of Greif and Tadelis (2010) are static in the sense that they do not intrinsically respond to coercion. The paper also contributes by introducing an endogenous dimension of

power: an insurrection constraint on the use of coercion that is dependent on the prevalence of non-state morality.

The paper also relates to the literature on ethnic and political violence, especially the understanding of use of force as a root cause of counter-mobilization in the form of a strategic response (Acemoglu and Wolitzky, 2014), or increased saliency of identity because of group conflict (Sambanis and Shayo, 2013). Acemoglu and Wolitzky (2014) focus on the informational aspects of group conflicts that lead to hostile actions followed by hostile reactions. They develop a dynamic Bayesian game of sequential aggressive or conciliatory actions between groups, where the driving static is whether agents interpret the hostile actions of opposing groups as the actions of a fundamentally aggressive type, or the actions of a non-aggressive type retaliating. They consider the informational aspect of group conflict, while this paper analyzes group conflict driven by an intrinsic reaction. Sambanis and Shayo (2013) build a formal model endogenizing the process of identification with an ethnic group. They allow for identification on multiple levels and focus on a social identity equilibrium between groups, where saliency determines the level of identification. Both these papers consider group relations and their internal dynamics, while in the model presented here, the agency lies in the state authority and population responses to the level of coercion.

Furthermore, the paper relates to a growing new literature on state legitimacy. The role of the state in nation building is formally analyzed in the economics literature by Alesina and Spolaore (2003) and Alesina and Reich (2013), while Greif and Rubin (2014) illustrates the need for independent agencies to provide legitimacy to the state. Greif and Rubin (2014) consider how the English Crown's breach with the Catholic

Church created a need for a new external agent to legitimize the king's power, thus increasing the need for approval of an independent agent, such as parliament. Johnson and Koyama (2013) investigate the relationship between the legitimacy gained by aligning the state with a specific religious belief rather than a compromise between several, and the economic cost of enforcing that belief. Whereas these papers focus on different sources of legitimacy and alignment between state and morality, this paper focuses on the use of force, its military constraints, and intrinsic reaction to the use of force to change moralities.

Finally, this paper relates to recent work by Acemoglu and Robinson (2017) in developing dynamic models where path dependency arises in societal outcomes. Acemoglu and Robinson (2017) develop a model of dynamic contest for power, where the state and society sequentially make costly investments in conflict capital. They find path dependency in the power of the state because of the discouragement effect of competitions: the interaction between incentives to invest and economies of scale in capital. This mechanism leads to a dynamic where either state and society invest in conflict capital to be thus equally matched in power, or, one of the parties ceases to invest and has no power. While Acemoglu and Robinson (2017) focus on conflict capital, this paper models dynamics of available labour for conflict, i.e., sizes of morality groups, when the coercion level changes the size of groups that determine the ability to coerce without having an insurrection. Path dependency arises as non-linearities in response to coercion, and the initial size of morality groups determine which coercion levels can be implemented by iteratively changing the coercion level.

2. A basic 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 morality of the parents. First, the basics of the model and the mechanisms of socialization are developed. We then develop assumptions regarding the parents' utility function and derive its implications. All the results here mirror the results from Bisin and Verdier (2001).

2.1. 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 moralities, $m \in \{a, b\}$. Moralities are mutually exclusive; a portion q_t of the parent population holds morality a at time t, while $1 - q_t$ holds morality b. Moralities 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 morality of the parent; and $1 - \tau^m$ that parental socialization fails, in which case the child is obliquely socialized and the offspring will adopt either morality a or morality b with a probability equal to the morality's prevalence in the population. A child who internalizes morality m is referred to as an m morality child. Let P^{mn} be the probability that an individual of morality m has an n morality offspring.

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

$$P^{bb} = \tau^b + (1 - \tau^b)(1 - q_t) \ , \ P^{ba} = (1 - \tau^b)q_t \tag{2}$$

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

$$q_{t+1} = q_t P^{aa} + (1 - q_t)P^{ba} = q_t + q_t(1 - q_t)(\tau^a - \tau^b) \tag{3}$$

From (3) it follows that the change in the share of morality a individuals is given by $q_t(1-q_t)(\tau^a-\tau^b)$: the difference in the probability of successful parental socialization times the product of the share of moralities.

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 morality parent having an n morality child be denoted u_n^m, then using (1) and (2), we attain the following utility function U^m for parents.

$$U^a = [\tau^a + (1 - \tau^a)q_t]u_a^a + (1 - \tau^a)(1 - q_t)u_a^b - H(\tau^a) \tag{4}$$

$$U^b = [\tau^b + (1 - \tau^b)(1 - q_t)]u_b^b + (1 - \tau^b)q_t u_b^a - H(\tau^b) \tag{5}$$

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

Assumption 1. Own morality preference Parents prefer their child to have the same morality as themselves: $u_a^a - u_b^a > 0,\ u_b^b - u_a^b > 0$.

Second, the utility loss of having a child internalize an opposing morality 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 morality and $\underline{u}$ as the utility derived from having

the opposing morality, we can write the following assumption.

Assumption 2. Symmetric utility loss of opposing morality Parents of a and b morality have symmetric utility loss in having children of opposing morality: $u_a^a - u_b^a = u_b^b - u_a^b = \overline{u} - \underline{u} = \Delta u$.

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

Assumption 3. 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 3 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 3 state that the marginal cost approaches infinity as the probability of having a child successfully socialized into the preferred morality 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 morality 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 morality in mixed morality populations. We can now derive the optimal levels of τ^m

from (4) and (5), which are given by the first order conditions (FOCs).

$$H'(\tau^a) = (1 - q_t)\Delta u,\ H'(\tau^b) = q_t \Delta u \tag{6}$$

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 (6), we can establish the following lemma.

Lemma 1. The smallest morality group always invests more in parental socialization: $\tau_t^b \leq \tau_t^a$ if and only if $q_t \leq (1 - q_t)$.

As the benefit of having a child with the parents' morality 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 morality differs because of different group size. Minority parents have a higher probability of their child internalizing the majority morality 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 (3) that for $q_t = q_{t+1}$ to be fulfilled, $q_t(1 - q_t)(\tau^a - \tau^b) = 0$ must hold. This is the case for $q_t = q_{t+1} = 0$, $q_t = q_{t+1} = 1$, i.e., single morality 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 morality, and oblique socialization will lead to the preferred morality of the parent. The single morality equilibrium is, however, unstable in the event of external shocks; if one parent of another morality enters the population,

this parent would choose a very high investment in parental socialization because the probability of the child adopting the desired morality in the case of oblique socialization would be very low. This would be repeated for future generations and consequently, the prevalence of introduced morality of the minority would grow until the unique interior $q^* = \frac{1}{2}$ is reached.

Lemma 2. 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. If the population is out of the SSE, the share of minority morality individuals will grow with time as the smaller morality 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 $\tau^a = \tau^b$, leads to an asymmetrical, i.e., $q^* \neq \frac{1}{2}$, stable SSE.[6]

3. Legitimacy maximizing given coercion resentment and an insurrection constraint

We now extend the model to include an authority that can issue a penalty, referred to as coercion, for adhering to non-state morality. 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

[6] 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 moralities themselves. The following analysis generalizes for asymmetrical preferences.

imposed. To focus on the implications of coercion resentment, we follow Greif and Tadelis (2010) in assuming that the authority can impose coercion at zero cost. The results can be extended trivially to a model where coercion is costly to the authority.

3.1. The extended model

There is an authority β controlling the state, where a state is defined as a monopoly on the employment of coercion, π, within the territory where the population is situated. This authority builds its legitimacy on b morality and wishes to maximize its prevalence by imposing coercion for adhering to a morality. The utility maximization problem of the authority, U^{β}, is as follows.

$$\max_{\pi} U^{\beta} = \min_{\pi} \; q^{*}(\pi) \tag{7}$$

To maximize the prevalence of b morality, the authority sets the level of coercion π for adhering to morality a. The level of coercion is assumed to be unbounded; π is defined over the domain $\pi \in [0, \infty)$, but we assume that the authority is bound by an upper feasibility constraint π_{max} on the level of coercion it can impose. Hence, we restrict our analysis within the feasible interval $\pi \in [0, \pi_{max}]$.[7]

The coercion level can be interpreted as ranging from low, such as social sanctions or issuance of fines for having morality a, to high, such as criminal penalties, and the maximum feasible level, π_{max}, is referred to as a gunpoint threat. Including the level of coercion, π, and resentment towards b morality caused by coercion, $C(\pi)$, in the utility function of the parents of a and b morality respectively, yields the following.

$$U^{a} = [\tau^{a} + (1 - \tau^{a})q_t](\overline{u} - \pi) + (1 - \tau^{a})(1 - q_t)(\underline{u} - C(\pi)) - H(\tau^{a}) \tag{8}$$

[7] Feasibility might reflect either technological constraints in terms of what can be implemented, or an upper limit in terms of what the state apparatus will impose.

$$U^b = [\tau^b + (1-\tau^b)(1-q_t)]\overline{u} + (1-\tau^b)q_t(\underline{u} - \pi) - H(\tau^b) \tag{9}$$

These two utility functions capture the two following assumptions of how agents respond to coercion.[8] First, we assume that the utility of having an a morality child is lower when there is coercion.

Assumption 4. Parental empathy for coercion The utility of having an a morality child is $(u_a^m - \pi)$.

Second, we assume coercion resentment, imposing coercion invokes a negative intrinsic reaction among the a non-state morality parents; i.e., they will have lower utility in having a b morality child.

Assumption 5. Coercion resentment The utility to an a parent of having a b morality child is $(u_b^a - C(\pi))$.

Note that the reduced utility, from introducing coercion, of an a morality child is the same for parents of both moralities. This reflects the fact that a morality children face an extrinsic cost, while the coercion resentment cost of having a b morality child, an intrinsic loss of utility, is limited to non-state morality parents. We now derive

[8] At $q_t \in \{0, 1\}$ parents will be indifferent between successful parental socialization or oblique socialization. As the cost of investing some infinitesimal amount in socialization or investing nothing $\tau^m = 0$ is equal, they will be indifferent between these two outcomes. We impose that $\tau^m = 0$ for $q_t \in \{0, 1\}$.

the optimal levels of τ^m from (8) and (9), which are given by the FOCs.

$$(1 - q_t)(\Delta u - (\pi - C(\pi))) = H'(\tau^a) \tag{10}$$

$$q_t(\Delta u + \pi) = H'(\tau^b) \tag{11}$$

Comparing (10) and (11) with the FOCs in the baseline model, (6), we see that the b morality parent increases socialization investment as π incurs a more severe utility loss if parental socialization fails and the child obliquely internalizes a morality. For a morality parents, coercion introduces two opposing effects: they have an incentive to reduce their investment, as having an a morality child becomes less extrinsically beneficial; and an incentive to increase investment, as b morality becomes less intrinsically beneficial as a result of coercion resentment. Without assuming a functional form on the coercion resentment function, we cannot say which effects dominate at which coercion levels.

As developed in the basic model of socialization, we see from (3) that a necessary condition for a stable interior SSE level is equal levels of investment, τ^m, in parental socialization of a and b morality. If parents invest equally in socialization, they have equal marginal costs: $H'(\tau^a) = H'(\tau^b)$. Hence, we can use (10) and (11) and establish the following lemma.

Lemma 3. 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 $q^*(\pi) = \frac{\Delta u - \pi + C(\pi)}{2\Delta u + C(\pi)}$.

This result is a basic extension of Proposition 1 in Greif and Tadelis (2010). The stationarity properties of (3) imply that the population will always converge to its SSE value. The Inada assumptions on investment in socialization imply positive

investment in socialization for a π corresponding to an internal SSE, hence the population will not reach any exterior solution in the convergence process as long as the SSE is internal. We define an initial interior SSE as some $q^*(\pi_0) \in (0,1)$, where π_0 is some initial coercion level $\pi_0 \in [0, \pi_{max}] : q^*(\pi_0) \in (0,1)$ and establish this as the following lemma.

Lemma 4. Imposing a coercion level π' corresponding to an internal SSE $q^*(\pi') \in (0,1)$ from an initial interior SSE $q^*(\pi_0)$, will make q converge to $q^*(\pi')$.

This result is a basic extension of Proposition 1 and Proposition 2 in Bisin and Verdier (2001). To illustrate the dynamics of the model, let us assume that at time t the coercion level is $\underline{\pi}$ and the population is in an interior SSE with $q^*(\underline{\pi})$. Assume that the value of π changes at $t+1$ to $\overline{\pi}$, where $\overline{\pi} > \underline{\pi}$, and that the net effect of coercion for a morality parents, $(\pi - C(\pi))$, is sufficiently increasing in the interval $[\underline{\pi}, \overline{\pi}]$ such that $q^{*\prime}(\pi) < 0$.[9] At $t+1$, q remains unchanged but investment in socialization changes; the a parents will now invest less in socialization as they have a net lower utility in having a morality children, while the b parents will invest more in socialization as the outcome of unsuccessful parental socialization, having an a morality 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, a morality parents will face a higher probability of their offspring having b morality through oblique socialization and will consequently in-

[9]This implies the average $C'(\pi)$ is less than two in the interval, as can be seen from (31) in Appendix 2.

crease their parental socialization. The level of the minority morality q_t will converge towards $q^*(\overline{\pi})$ until the SSE condition from (3) of $\tau^a = \tau^b$, i.e., equal investment in parental socialization, is restored at the SSE with $q^*(\overline{\pi})$.

Imposing a coercion level that does not correspond to an internal SSE must imply a value of π such that one of the morality groups will cease to invest in socialization. This will lead to $q^*(\pi)$ reaching the external SSE without the morality group that ceases to socialize their children within one generation.

We now discuss the coercion resentment function. The 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, i.e., assumed to be simply π. Thus, discussion of the net effect of coercion for a morality parents can be centred around the coercion resentment function, $C(\pi)$. First, some fairly unrestrictive functional form assumptions are made of $C(\pi)$:

$$C(\pi) \text{ is a function of the } C^2 \text{ class, it is } C(0) \geq 0 \text{ and it has } C'(\pi) > 0, \text{ over the domain } [0, \pi_{max}]. \tag{12}$$

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

The coercion resentment function is assumed to be S-shaped.[10] The convex part of the coercion function captures the idea that there is an increasing marginal emotional response to an increase in π for initial levels of coercion. As the authority

[10] As shown in Appendix 2, linear, convex or concave coercion resentment functions have trivial and unique optimums.

increases π, it goes from being perceived as representative of b morality individuals, which favours and endorses b morality, to being perceived as an enemy of a morality individuals, with aggressive intentions of reducing the prevalence of a morality individuals. The concave interval means that the change in this response becomes negative beyond some point; as the intentions of the authority become clear, higher levels of coercion cause a smaller increase in resentment. We define a point $\hat{\pi}$ in the open interval, $\hat{\pi} \in (0, \pi_{max})$ and assume that:

$$C''(\pi) = \begin{cases} > 0 & \text{for} \quad \pi \in [0, \hat{\pi}) \\ = 0 & \text{for} \quad \hat{\pi} \\ < 0 & \text{for} \quad \pi \in (\hat{\pi}, \pi_{max}]. \end{cases} \tag{13}$$

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

Assumption 6. Varying 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, $\pi' \in (0, \pi_{max})$; $C'(\pi') > 2$.

We define a coercion level π' as marginally effective if $q^{*\prime}(\pi') < 0$. Using the assumptions of $C(\pi)$ in (12), (13) and Assumption 6, we can develop the following lemma on the overall effect of coercion.

Lemma 5. Coercion is marginally effective at the beginning and at the end of $[0, \pi_{max}]$, and there is at least one level of coercion, $\hat{\pi}$, that is strictly marginally ineffective: $q^{*\prime}(\hat{\pi}) > 0$.

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, i.e., 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.

Whenever $q^*(\pi)$ is strictly positive for all $\pi \in [0, \pi_{max}]$, the assumptions that $C(\pi)$ is monotonically increasing in π and has a continuous second derivative, imply that $q^*(\pi)$ will always have a unique infimum in the open convex part of $C(\pi)$, $(0, \hat{\pi})$ where $q^{*\prime}(\pi) = 0$. We denote the coercion level giving this infimum as $\pi_{\underline{q}} \in (0, \hat{\pi})$, and refer to it as a nonconfrontational 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:

$$\pi^e_{\underline{q}} \text{ is defined as a coercion level such that } \pi_{\underline{q}} < \pi^e_{\underline{q}} \text{ and } q^*(\pi^e_{\underline{q}}) \equiv q^*(\pi_{\underline{q}}). \qquad (14)$$

$\pi^e_{\underline{q}}$ will only be defined for functional forms where $C(\pi)$ is sufficiently concave in $(\hat{\pi}, \pi_{max}]$. There will always be a unique supremum value of $q^*(\pi)$ in the concave part of $C(\pi)$; we denote this level as $\pi_{\overline{q}} \in (\hat{\pi}, \pi_{max}]$. Applying Lemma 5 to the effect of coercion, the assumptions placed on $C(\pi)$ in (12) and (13) and Lemmas 3 and 4 on the interior SSE, we can develop Lemma 6 on the functional form of $q^*(\pi)$.

Lemma 6. $q^*(\pi)$ 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, poten-

tially 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 $q^*(0) = \frac{1}{2}$. [11]

The properties of $q^*(\pi)$ are dependent on the size of utility loss for parents from having children with differing morality, 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 non-zero minimum $q^*(\pi)$ value. Class III) is qualitatively similar to a linear coercion resentment function, i.e., $C(\pi) = K_0 + K_1\pi$, with $K_1 > 2$, while II) is qualitatively non-convex. Figure 1. illustrates the three possible classes of $q^*(\pi)$.

3.2. Analysis

We now analyze the optimal level of coercion for an authority minimizing the share of individuals with *a* morality in SSE, $q^*(\pi)$. We assume that the authority chooses π from an initial interior SSE, $q^*(\pi_0)$. First, we discuss the model without any constraint on the use of coercion within $[0, \pi_{max}]$. Second, we discuss the optimal coercion levels under an exogenously given constraint on coercion, $\rho \in (0, \pi_{max})$. Finally, we discuss the model under an endogenously given constraint $\rho(q^*(\pi)) \in [0, \pi_{max}]$.

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

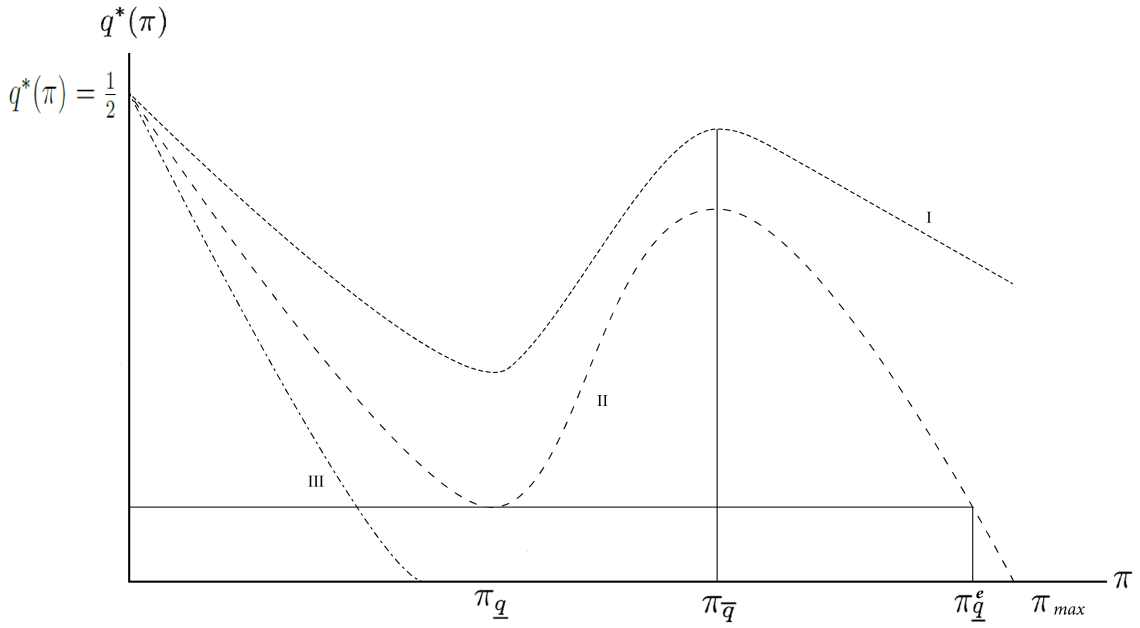


Figure 1: Three examples of $q^*(\pi)$ from $\pi_0 = 0$, constructed using $C(\pi) = \tan^{-1}\pi$.

The no constraint analysis is done to analyze how different functional forms on $C(\pi)$ within $[0, \pi_{max}]$ change the optimum. The constraints added within the $[0, \pi_{max}]$ interval are strategic constraints that are used to show how a change in a constraint within a given interval changes the optimal level of coercion. As the initial coercion level π_0 will not influence the optimal choice of the authority for the no or exogenous constraint cases, it is omitted from the analysis of these cases and only introduced for the endogenous constraint analysis.

Note that in this analysis we do not restrict the optimal choice of π to lead to an interior solution; if the authority can impose a π' that corresponds to the exterior solution, $q^*(\pi') = 0$, it will do so.[12]

No constraint

As established in Lemma 4 and subsequent discussion, an authority in $q^*(\pi_0)$ can choose any feasible π' and will always converge to the corresponding $q^*(\pi') \in [0, 1]$. Applying Lemma 3 and Lemma 4 to the utility function in (7), the maximization problem for an authority is given by the following.

$$\max_{\pi} U^{\beta} = \min_{\pi}[\frac{\Delta u - \pi + C(\pi)}{2\Delta u + C(\pi)}] = \min_{\pi} q^*(\pi) \tag{15}$$

The optimal coercion level with no constraints on the use of coercion, π^{NC}, is de-

[12] The utility function in (7) implies that whenever a set, i.e., multiple, (π'', π''') corresponds to $q^*(\pi) = 0$, the authority will be indifferent as to which $\pi \in (\pi'', \pi''')$ to impose; by institution we refer to the imposed level as the lowest π that attains $q^*(\pi) = 0$. Once the population is in an exterior SSE, q^* will no longer be a function of π and the model is silent on which π the authority will impose.

termined by the properties of $q^*(\pi)$, given by $\{\Delta u, C(\pi)\}$. Consequently, the results follow directly from Lemma 6. For sufficiently strong coercion resentment and large Δu, class I) applies. For class I), the imposed level of coercion will be $\pi^{NC} = \pi_{\underline{q}}$, as $q^*(\pi_{\underline{q}})$ is strictly smaller than $q^*(\pi_{max})$ and will consequently be preferred by the authority. Hence, when coercion resentment is sufficiently strong, such that $q^*(\pi)$ is at its minimum for unconfrontational levels of coercion, the authority will not impose the maximum level of coercion, even when it is able to do so. The population will remain in an interior solution in the presence of both morality types with an unconfrontational level of coercion.

If class II) applies for $q^*(\pi)$, the optimal level of coercion will be $\pi^{NC} \in (\pi^e_{\underline{q}}, \pi_{max}]$ if any π' gives $q^*(\pi') = 0$. Otherwise, $\pi^{NC} = \pi_{max}$ will be the optimal level of coercion and the population will be stable at an interior solution at $q^*(\pi_{max})$. If class III) applies to $q^*(\pi)$, Δu is sufficiently low and coercion resentment is sufficiently weak so that the authority can impose a coercion level that is lower than the unconfrontational level, and attain $q^* = 0$. We sum up the no constraint analysis in Proposition 1 as follows.

Proposition 1. Coercion use under no constraint Let π' denote a level of coercion such that $q^*(\pi') = 0$. The optimal level of coercion under no constraint, π^{NC}, will be as follows for the different classes of $q^*(\pi)$.
I) $\pi^{NC} = \pi_{\underline{q}}$
II) $\pi^{NC} = \pi' \in (\pi^e_{\underline{q}}, \pi_{max})$, if no π' is defined then $\pi^{NC} = \pi_{max}$
III) $\pi^{NC} = \pi' < \pi_{\underline{q}}$, where π' is always defined.

Referring to an authority that imposes a coercion level strictly lower than its highest

implementable level towards a non-zero a morality group as exhibiting restraint, we can establish the Corollary of Proposition 1 as follows.

Corollary of Proposition 1. Restraints under no constraint An authority facing no constraint on coercion will only restrain its use of coercion when SSE $q^*(\pi)$ is of class I).

Exogenous constraint

Several factors external to the model can constrain the use of coercion by an authority: the authority might recognize constitutional legal rights, there might be institutionalized rights constraining what π the state apparatus can issue, or surpassing a coercion threshold might trigger an intervention by foreign powers. To analyze optimal 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. We assume an initial coercion level $\pi_0 \in [0, \rho]$ from which any $[q^*(0), q^*(\rho)]$ can be reached. In addition to determining the optimal π^{EC}, we also focus on when the constraint will be binding at the optimal coercion level.[13]

Including a constraint on the use of coercion leaves an authority with the follow-

[13]Binding constraints might change the level and the saliency of conflict between morality groups. Furthermore, it will presumably be easier to empirically observe binding constraints, such as explicit threats of intervention and emerging population movements, than unbinding and latent constraints that might be in the form of unrealized outcomes anticipated by an authority.

ing optimization problem.

$$\max_{\pi} U^{\beta} = \min_{\pi}[\frac{\Delta u - \pi + C(\pi)}{2\Delta u + C(\pi)}] \text{ s. t. } \pi \leq \rho \tag{16}$$

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, $\rho < \pi^{NC}$. Noting this, we can develop the following proposition on the optimal level of coercion, π^{EC}, for an authority facing a constraint on the use of coercion.

Proposition 2. Coercion use under an exogenous constraint 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 $\rho \in (\pi_{\underline{q}}, \pi^{e}_{\underline{q}})$.[14]

The result shows that constraints in $\rho \in (\pi_{\underline{q}}, \pi^{e}_{\underline{q}})$, an interval of coercion that an authority will always find undesirable to impose and referred to as the inefficient interval of coercion, lead to an imposed level of coercion $\pi^{EC} = \pi_{\underline{q}}$ with a constraint that is non-binding in the optimum: $\pi^{EC} < \rho$. Furthermore, if ρ changes from within $\rho' \in (\pi_{\underline{q}}, \pi^{e}_{\underline{q}})$ to some level $\rho'' > \pi^{e}_{\underline{q}}$, the coercion level will jump discontinuously from $\pi^{EC} = \pi_{\underline{q}}$ to $\pi^{EC} = \rho''$.[15]

Considering Lemma 6, we see that the inefficient interval is only defined for functional form II); we use this together with Proposition 2 to develop the following proposition.

[14] With $\rho = \pi^{e}_{\underline{q}}$ the authority is indifferent between imposing $\pi_{\underline{q}}$ and $\pi^{e}_{\underline{q}}$.

[15] Proposition 2 has relevant implications for the policy problem of an external agency setting a constraint ρ to limit an authority's use of coercion when $q^{*}(\pi)$, as discussed in Appendix 3.1.

Corollary of Proposition 2. Restraints under constraint An authority will restrain its use of coercion as a response to a constraint if and only if the $q^*(\pi)$ is of class II) and the constraint is in the inefficient interval of coercion: $\rho \in (\pi_{\underline{q}}, \pi^e_{\underline{q}})$.

Hence, a legitimacy-maximizing authority rationally restraining its level of coercion as a response to a constraint, must imply a non-convex response to coercion. In this model, it also implies that the imposed coercion level π^{EC} is equal to the unconfrontational level of coercion $\pi_{\underline{q}}$.

Endogenous insurrection constraints

We now analyze the model, assuming an endogenous insurrection constraint on the use of coercion, dependent on the initial prevalence of *a* morality, $q^*(\pi_0)$. For tractability, the insurrection constraint $\rho(\cdot)$ is assumed to be dependent on $q^*(\pi)$ rather than on q_t: $\rho(q^*(\pi))$.[16]

We first define the insurrection constraint, then we show that the solution to the static optimizing problem of setting π from an initial π_0 is not necessarily an equi-

[16] A description of which conditions are needed for equivalence between solving the authority's optimization problem constrained by an insurrection constraint dependent on the SSE $q^*(\pi)$, $\rho(q^*(\pi))$ or by a constraint dependent on q_t, is included in Appendix 4.1.

librium if the authority has the opportunity to reset π in the new SSE, $q^*(\pi)$.[17] To address this issue, we develop a formal definition of the set of implementable coercion levels, $\mathbf{S}_{\pi_0}$, for an authority with a given initial condition π_0. To find which of the implementable levels of coercion will lead to an equilibrium outcome, we develop a notion of dynamically stable equilibria, characterized by the authority not having an incentive to change π^{IC} if this was the initial coercion level: $\pi_0 = \pi^{IC}$.[18] Finally, we investigate which coercion levels characterize dynamically stable equilibria and show how the model may display path dependence; i.e., different initial conditions may give different equilibria.

The endogenous insurrection constraint is defined as the highest coercion level for which the minority has negative excepted utility of committing an insurrection. The insurrection constraint function $\rho(q^*(\pi_0))$ defines the maximal coercion level that can be implemented for some initial state $q^*(\pi_0)$ without the a morality committing an insurrection. Note that there is no explicit link between the insurrection decision and coercion resentment; the private decision processes of how much to invest in socialization may be very different to the public decision processes for a morality group to commit an insurrection. There is no specified outcome for an insurrection; because we assume the authority sets π in order to avoid an insurrection, we implicitly assume that the authority must find the insurrection outcome to be worse

[17]Note that because Lemma 5 implies convergence from an internal to another internal SSE, we cannot say that the authority can reset π once q^t reaches $q^*(\pi)$. The issue can be solved by assuming that the authority can reset π once q^t is within some infinitesimal interval ϵ of $q^*(\pi)$. This is, however, not necessary to address which stable equilibria exist and are reachable within $t \in [0, \infty)$, which is the subject of this model, hence we omit this complication.

[18] A standard definition of stability in dynamic games (Petrosyan, 2016).

than being able to reset π, satisfying the constraint. Implicitly, we also assume that the minority might avoid or reduce coercion given a successful insurrection. We first impose the following assumption on the insurrection constraint.

Assumption 7. Monotonically increasing insurrection constraint The insurrection constraint $\rho(q^*(\pi))$ is a continuous mapping from $q^*(\pi) \in (0,1)$ to $[0, \pi_{max}]$. It is monotonically decreasing in $q^*(\pi)$ and has a continuous first derivative.[19]

We insert the endogenous insurrection constraint into (15) to attain the authority's static optimization problem with an endogenous insurrection constraint in the following.

$$\max_{\pi} U^{\beta} = \min_{\pi}[\frac{\Delta u - \pi + C(\pi)}{2\Delta u + C(\pi)}] = \min_{\pi} q^*(\pi) \text{ s.t. } \pi \leq \rho(q^*(\pi_0)) \tag{17}$$

Unless the authority can set π only once, and is unable to subsequently readjust its π, the solution to (17) is not necessarily a dynamically stable equilibrium. As the insurrection constraint is dependent on $q^*(\pi)$, choosing 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 (17) may be dynamically unstable in the sense that the authority may have an incentive to set a new $\pi'' > \pi'$ in order to attain a lower SSE, $q^*(\pi'') < q^*(\pi')$.

[19] We discuss interpretations of the insurrection constraint in Appendix 3.2.

To find the dynamically stable coercion level an authority *will* implement, we first develop a formal notion of which coercion levels an authority *can* implement if it has the opportunity to reset π an infinite number of times, $\mathbf{S}_{\pi_0}$. We first define the set of sustainable coercion levels, $\mathbf{S}_{\Pi} \equiv \{\pi : \pi \leq \rho(q^*(\pi))\}$: 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 $\pi \in \mathbf{S}_{\Pi}$ are *implementable* from a given initial condition π_0, $\mathbf{S}_{\pi_0}$ is a subset of the set of sustainable levels $\mathbf{S}_{\pi_0} \subseteq \mathbf{S}_{\Pi}$. We formally define the set of implementable coercion levels, $\mathbf{S}_{\pi_0}$, from an initial condition π_0 as follows.

Definition of the set of implementable coercion levels: A coercion level π' is in the set of implementable coercion levels $\mathbf{S}_{\pi_0}$ 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; $\pi_n \in \mathbf{S}_{\Pi}$ for all $n = 1, 2...N$.

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.[20][21]

To further study $\mathbf{S}_{\pi_0}$, we develop the composite function $\hat{\rho}(\pi) \equiv \rho(\cdot) \circ q^*(\pi)$ as

[20] By assumption, the initial condition corresponds to a sustainable level of coercion, $\pi_0 \leq \rho(q^*(\pi_0))$.

[21] Implicitly, we here assume that π can only be reset once a $q^*(\pi)$ is reached. A discussion of the set of implementable coercion levels where the authority can reset π at any t is included in Appendix 4.2.

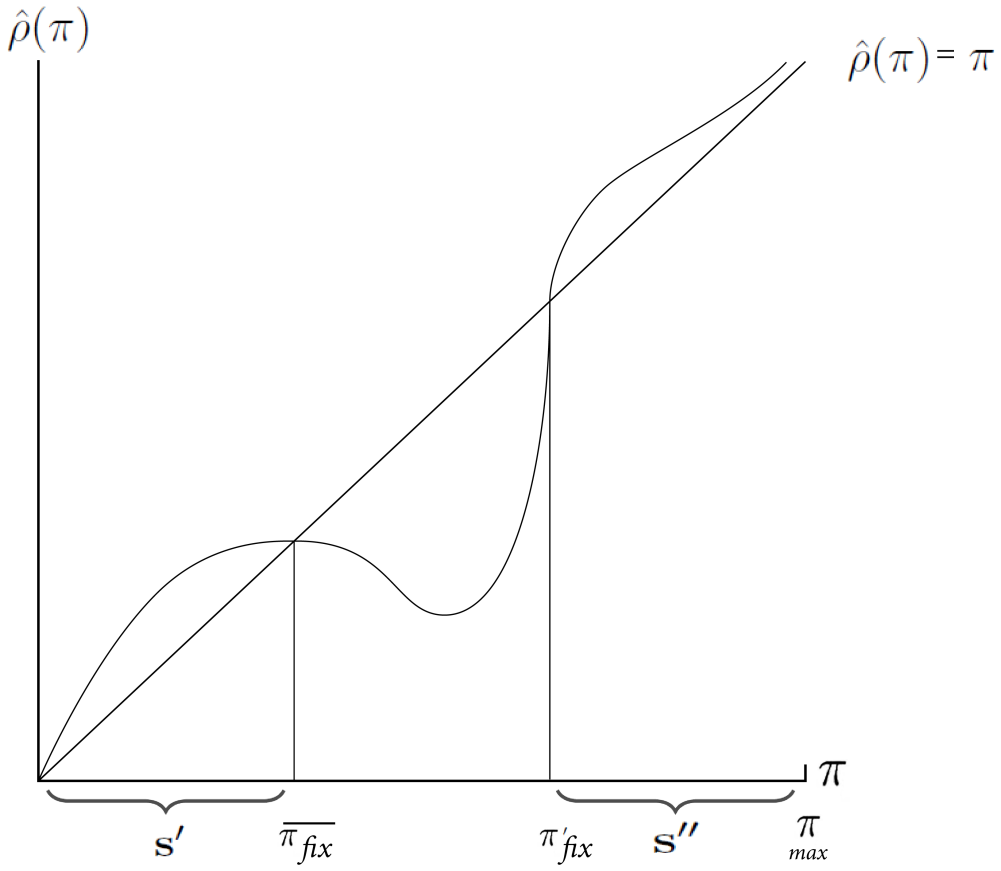


Figure 2: The curved line is an example of $\hat{\rho}(\pi)$ while the 45-degree line is the fix-point-line. Any $\pi_0 \leq \overline{\pi}_{fix}$ will give $\mathbf{S}_{\pi_0} = \mathbf{s}' = [0, \overline{\pi}_{fix}]$, while a $\pi_0 \geq \pi'_{fix}$ will give $\mathbf{S}_{\pi_0} = \{\mathbf{s}', \mathbf{s}''\} = \{[0, \overline{\pi}_{fix}], [\pi'_{fix}, \pi_{max}]\}$.

the composite of the insurrection constraint $\rho(q^*(\pi))$ and $q^*(\pi) \in (0,1)$, i.e., a value of $\hat{\rho}(\pi')$ is the insurrection constraint at the SSE corresponding to π': $q^*(\pi')$.[22] The functional form of $\hat{\rho}(\pi)$ will determine the properties of $\mathbf{S}_{\pi_0}$ and will be determined by the form of $q^*(\pi)$ in conjunction with the form of the insurrection constraint $\rho(\cdot)$. As $\rho'(q^*(\pi))$ is assumed to be monotonically increasing in $q^*(\pi)$, the functions $\hat{\rho}(\pi)$ and $q^*(\pi)$ will always have derivatives of equal sign.[23] Plotting an illustration of the $\hat{\rho}(\pi)$ function exemplified by a convex $\rho(\cdot)$ function is done in Figure 2.

Not all implementable coercion levels are sustainable. If a coercion level π' increases the SSE from $q^*(\pi_0)$, it will decrease the insurrection threshold of the a morality group and might lead to an insurrection at $q^*(\pi')$. Hence, a coercion level π' might be implementable, $\hat{\rho}(\pi_0) > \pi'$, lower than an authority's initial condition, $\pi' < \pi_0$, but still be unsustainably *low*. A coercion level may also be implementable but be unsustainably *high*. If an authority was to implement an implementable but unsustainable coercion level π', the state would remain stable at the time of implementation, but have an insurrection once q converges to its $q^*(\pi')$ level, at which the size of the a morality group is large enough as to choose to commit an insurrection at π'. Whenever a coercion level is implementable, $\hat{\rho}(\pi_0) > \pi'$, but also unsustainable, $\hat{\rho}(\pi') < \pi'$, it implies that $q^*(\pi)$ is strictly increasing in the interval between π_0 and π'. Hence, an authority minimizing $q^*(\pi)$ will never implement an unsustainably low π in a path towards an π^{IC}, as an unsustainable level of coercion must imply

[22] As $\rho(\cdot)$ is a continuous function mapping $q^*(\pi) \in (0,1)$ to $\pi \in [0, \pi_{max}]$, and the function $q^*(\pi) \in (0,1)$ is a continuous function mapping $[0, \pi_{max}]$ to $q^*(\pi) \in (0,1)$, the composite of the two, $\hat{\rho}(\pi)$ is a continuous function mapping $[0, \pi_{max}]$ to $[0, \pi_{max}]$.

[23] This trivially holds because $\hat{\rho}(\pi) \equiv \rho(q^*(\pi))$ and Assumption 7 statement $\rho(\cdot)' > 0$ for all $q^*(\pi) \in (0,1)$ implies that when $q^*(\pi)' > 0$ then $\hat{\rho}(\cdot)' > 0$.

imposing a π increasing the SSE.[24]

The non-linearities in $q^*(\pi)$ mean there might exist unsustainable levels of coercion π' between the upper and the lower bound of a $\mathbf{S}_{\pi_0}$. We define any subset of $\mathbf{S}_\Pi$ that is not a union of two disjoint non-empty open sets, as $\mathbf{s}$.[25] As $\hat{\rho}(\pi)$ may cross the fix-point-line multiple times, there may be multiple $\mathbf{s}$ sets separated by unsustainable π in a $\mathbf{S}_{\pi_0}$. As $\rho(q^*(\pi))$ is strictly decreasing in $q^*(\pi)$, an authority with $\pi_0 \in \mathbf{s}'$ will always be able to set $\pi' = \hat{\rho}(\pi_0)$ until it reaches an end point of $\mathbf{s}'$. However, because of non-linearities in $q^*(\pi)$ this is not necessarily true for $\mathbf{S}_{\pi_0}$; i.e., more sophisticated programs of changes in the coercion level might be needed.

Authorities with composite insurrection constraints such that zero is not a sustainable coercion level, $0 \notin \mathbf{S}_\Pi$, are defined to be strongly coercion reliant: they will be reliant on strictly positive levels of coercion to sustain their state, and impose $\pi > 0$ without any inherent incentive to minimize $q^*(\pi)$.[26] We establish the following as a formal definition of coercion reliance.

Definition of weak and strong coercion reliance: An authority is defined as *strongly coercion reliant* whenever $\hat{\rho}(0) < 0$ and *weakly coercion reliant* whenever there exists unsustainable levels of coercion π' that are lower than the initial condi-

[24] An example of an authority that could implement an unsustainably low π is one that seeks to gradually reduce π and knows the direction of $q^{*\prime}(\pi)$ but is uncertain of the magnitude.

[25] This definition is adopted from Mendelson (1975).

[26] Assuming that a group with superior coercive capability will always commit an insurrection, an authority with coercion reliance $0 \notin \mathbf{S}_{\pi_0}$ implies that the minority has superior coercive capability at $q^*(0) = \frac{1}{2}$. Note that, in a more general framework, this might occur whenever $q^*(0) > \frac{1}{2}$.

tion: $\pi' < \pi_0 : \pi' \notin \mathbf{S}_{\pi_0}$.

As $\rho(\cdot)$ is homogenously decreasing in $q^*(\pi)$, a strictly marginally inefficient interval of coercion between intervals of marginally effective π, such as $(\pi_{\underline{q}}, \pi_{\overline{q}})$, must exist for weak coercion reliance to occur independently of strong coercion reliance.[27] We establish this as Proposition 3 below.

Proposition 3. Weak coercion reliance and coercion inefficiency Weak coercion reliance occurring without strong coercion reliance implies a strictly marginally inefficient interval of coercion between intervals of marginal effective coercion levels.

The bounds of $\mathbf{S}_{\pi_0}$ will either be a coercion level at which the insurrection constraint binds the $q^*(\pi')$ to which it has converged; i.e., at the line of fix-points referred to as the fix-point-line, of $\hat{\rho}(\cdot)$ defined as $\pi_{fix} \equiv \{\pi : \hat{\rho}(\pi) = \pi\}$, or the bounds of $\mathbf{S}_{\pi_0}$ will be at the bounds of the $[0, \pi_{max}]$ interval. We define the different types of bounds on $\mathbf{S}_{\pi_0}$ as follows.

Definition of constraints on the set of implementable coercion levels: For any π_0, any upper or lower bounds of the set of implementable coercion levels from π_0, $\mathbf{S}_{\pi_0}$, will be either:
I. a *strategic* constraint, if the coercion level at the bound of $\mathbf{S}_{\pi_0}$ is a fix-point of the insurrection constraint, π_{fix}

[27] An interesting consequence of this is that under weak coercion reliance without strong coercion reliance, a gradual reduction of π towards 0 will lead to state failure, while a sudden change from π_0 to 0 will not.

II. a *feasibility* constraint, if the coercion level at the bound of $\mathbf{S}_{\pi_0}$ is π_{max}.

If the upper bound of $\mathbf{S}_{\pi_0}$ is a strategic constraint, defined as $\overline{\pi_{fix}} \equiv sup\{\mathbf{S}_{\pi_0}\}$, it must lie at a crossing of the fix-point-line by the composite insurrection function from above. An upper strategic constraint on $\mathbf{S}_{\pi_0}$, $\overline{\pi_{fix}}$, will be an attractor fix-point with $\hat{\rho}'(\pi) < 1$; the authority can increase π until it arrives at this level of coercion and will do so if this is the minimum of $\mathbf{Q}_{\pi_0}$. This property arises from the fact that $\hat{\rho}(\pi)$ is continuously defined, hence any coercion level π' in the open $\mathbf{S}_{\pi_0}$ set must satisfy $\hat{\rho}(\pi') < \pi'$, while a strategic constraint by definition is on fix-point-line where $\hat{\rho}(\pi_{fix}) = \pi_{fix}$. The opposite holds for a strategic constraint at the lower bound of $\mathbf{S}_{\pi_0}$, defined as $\underline{\pi_{fix}} \equiv inf\{\mathbf{S}_{\pi_0}\}$ because the insurrection function has a lower bound that is a crossing of the fix-point-line from below, and it holds that $\hat{\rho}'(\underline{\pi_{fix}}) > 1$. Hence, $\underline{\pi_{fix}}$ is a repeller fix-point; the response to the use of coercion changes the insurrection threshold sufficiently that the authority can increase π from $\underline{\pi_{fix}}$ to higher levels of coercion.

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

Definition of a dynamically stable equilibrium: A dynamically stable equilibrium is defined as a coercion level and an SSE $\{\pi^{IC}, q^*(\pi^{IC})\}$ such that π^{IC} is the

optimal coercion level if π^{IC} is equal to the initial condition π_0, $\pi_0 = \pi^{IC}$. [28]

This definition implies that a dynamically stable equilibrium $\{\pi^{IC}, q^*(\pi^{IC})\}$ must fulfil the following three conditions (conditions 2 and 3 follow from 1, but are included for completeness).

1. π^{IC} is the solution to the static optimization of the authority (17) when $\pi_0 = \pi^{IC}$.
2. Socialization investment for both morality groups is unchanged at $q^*(\pi^{IC})$, i.e., $q^*(\pi^{IC})$ is an SSE.
3. The a morality group does not choose to commit an insurrection at π^{IC}; i.e., the coercion level does not breach the insurrection constraint; $\pi^{IC} \leq \rho(q^*(\pi^{IC}))$.

Conditions 2 and 3 in the definition of a dynamically stable equilibrium are fulfilled for all $\pi \in \mathbf{S}_{\pi_0}$, while condition 1 is fulfilled by the π in $\mathbf{S}_{\pi_0}$ that maximizes U^β. Hence, we find π^{IC} by solving the following.

$$\pi^{IC} \equiv \{\underset{\pi}{\operatorname{argmax}}\, U^\beta = \min_{\pi \in \mathbf{S}_{\pi_0}} [\frac{\Delta u - \pi + C(\pi)}{2\Delta u + C(\pi)}]\} \tag{18}$$

Any authority that can infinitely reset π will always be at, or in a sequence $\{\pi_n\}_0^{N-1}$

[28] This is equilibrium in the infinitely repeated game where the authority first chooses π', whereupon a morality parents collectively choose whether to insurrect, and if not, parents of both morality groups choose their levels of socialization investment τ^a, τ^b. New generations of parents then keep setting 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. Note that the parents only have preferences for their child's morality, not for any later generation, which simplifies the strategic aspects of the game.

leading up to, a dynamically stable equilibrium $\{\pi^{IC}, q^*(\pi^{IC})\}$. The dynamically stable equilibrium π^{IC} will be unique corresponding to every feasible π_0, except for one special case.[29] The monotone derivative of the insurrection function, $\rho'(q^*(\pi)) < 0$, implies that at any iteration there can never be another $q^*(\pi)$ giving 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 dynamical equilibria is robust to the introduction of time preferences when coercion is costless and $\rho(q^*(\pi))$ is monotonically decreasing in $q^*(\pi)$.

Any dynamically stable equilibrium will either be at a strategic constraint, a feasibility constraint or a local minimizer of $q^*(\pi)$. We establish this as Proposition 4.

Proposition 4. Dynamically stable equilibria For any initial condition π_0, the dynamically stable equilibrium π^{IC} is a coercion level equal to either:

I. the unconfrontational level of coercion as an interior point of $\mathbf{S}_{\pi_0}$: $\pi^{IC} = \pi_{\underline{q}}$

II. a strategic constraint at the upper bound of $\mathbf{S}_{\pi_0}$: $\pi^{IC} = \overline{\pi_{fix}}$

[29]The only instance in which the minimum of $\mathbf{Q}_{\pi_0}$ does not correspond to a unique minimum π' is the case where $\{\pi_{\underline{q}}, \pi^e_{\underline{q}}\} \in \mathbf{S}_{\pi_0}$, $\hat{\rho}(\pi_{\underline{q}}) = \pi^e_{\underline{q}}$ and $\pi^e_{\underline{q}} = \overline{\pi_{fix}}$; then by definition $inf(\mathbf{Q}_{\pi_0}) = \{q^*(\pi_{\underline{q}}), q^*(\pi^e_{\underline{q}})\}$. Hence, $\pi^{IC} = \{\pi_{\underline{q}}, \pi^e_{\underline{q}}\}$ and the authority will be indifferent between imposing $\pi_{\underline{q}}$ or $\pi^e_{\underline{q}}$. In an application of the model, this issue can be solved by considering whether the relevant authority has other considerations that make coercion costly, in which case $\pi^{IC} = \pi_{\underline{q}}$, or beneficial, in which case $\pi^{IC} = \pi^e_{\underline{q}}$.

III. the upper feasibility constraint at the bound of $\mathbf{S}_{\Pi}$: $\pi^{IC} = \pi_{max}$. [30]

When $\{\Delta u, C(\pi)\}$ is sufficiently low such that $q^*(\pi)$ is of class III), π^{IC} will either be equal to a strategic constraint $\overline{\pi_{fix}} \in [0, \pi_{\underline{q}})$ or the lowest π' attaining $q^*(\pi) = 0$, depending on whether the composite insurrection function is such that π' is implementable from π_0. For $q^*(\pi)$ of class I), the dynamically stable π^{IC} is equal to $\pi_{\underline{q}}$ if this is implementable, and equal to some strategic constraint $\overline{\pi_{fix}} > \pi_{\underline{q}}$ if not. For $q^*(\pi)$ of class II), authorities will end up in stable gunpoint equilibria with two morality populations if π_{max} is implementable and $0 < q^*(\pi_{max})$. If there exists an implementable π' such that $q^*(\pi') = 0$, then the population will approach single morality equilibrium. If this is not the case, then either π^{IC} 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 $\overline{\pi_{fix}} \in [0, \pi_{\underline{q}})$ or below $\overline{\pi_{fix}} \in (\pi^e_{\underline{q}}, \pi_{max})$, the open interval of inefficient coercion levels, $(\pi_{\underline{q}}, \pi^e_{\underline{q}})$.

Equal to the model under a exogenous constraints, the equilibrium $\pi^{IC} = \pi_{\underline{q}}$ is the only one where the authority restrains its use of coercion in equilibrium; it imposes a coercion level strictly lower than the highest implementable coercion level. The equilibrium $\pi^{IC} = \overline{\pi_{fix}}$ is given by $\hat{\rho}(\pi)$ and implies a coercion level at a binding insurrection constraint. Finally, the equilibrium where $\pi^{IC} = \pi_{max}$ can be under-

[30] Note that because equilibrium is defined as mutual reinforcing behaviour *determining* an outcome, the equilibrium with a coercion level similar to the feasibility constraint at π_{max} could be omitted, as it is the feasibility constraint that determines $\pi' = \pi^{IC}$ and is not set at any threshold where π changes a strategic choice made by the a morality group to commit an insurrection.

stood 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 π_{max}.

When a composite insurrection function is such that different initial conditions, π_0, generate different $\mathbf{S}_{\pi_0}$ sets, the model will have path dependency; different initial conditions will give different dynamically stable equilibria, π^{IC}. For a composite insurrection function $\hat{\rho}(\pi)$ with two crossings at the fix-point-line, such as the one described in Figure 1, $\mathbf{S}_\Pi$ will consist of two disjoint subsets, $\mathbf{s}$, one lower $\mathbf{s}'$ and one upper $\mathbf{s}''$. Assume that $\overline{\mathbf{S}}_{\overline{\pi_0}}(\pi) = \{\mathbf{s}'', \mathbf{s}'\}$, then $\underline{\pi_0} \in \mathbf{s}'$ and $\overline{\pi_0} \in \mathbf{s}''$ will produce different sets of implementable coercion levels, depending on whether there is any way of implementing the minimal π' in the upper subset $\mathbf{s}''$ from the lower $\mathbf{s}'$; that is, if $\rho(inf\{\mathbf{Q}_{\underline{\pi_0}}\}) < inf\{\mathbf{s}''\}$.

Generally, lower subsets of $\mathbf{S}_\Pi$, such as $\mathbf{s}'$, will always be contained in higher sets of implementable coercion levels, such as $\overline{\mathbf{S}}_{\overline{\pi_0}}$. This arises because an authority never instantaneously triggers an insurrection by lowering the coercion level. Only when coercion levels lower than the initial condition are unsustainable, can reductions in coercion lead to an insurrection. Hence, path dependency only goes from low to high levels of coercion; an authority can be restricted to impose a lower level of coercion than it would otherwise be able to impose, but cannot be restricted to impose higher levels of coercion than it otherwise could because of historical factors. In other words, the lower bound of $\mathbf{S}_\Pi$ is the lower bound of any $\mathbf{S}_{\pi_0}$. We conclude our formal analysis by developing path dependency as a proposition.

Proposition 5. Path dependency 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, $\overline{\mathbf{S}}_{\overline{\pi}_0} \triangle \underline{\mathbf{S}}_{\underline{\pi}_0} \neq \emptyset$, will different initial conditions $\pi_0 = \overline{\pi}_0$ and $\pi_0 = \underline{\pi}_0$ give different dynamically stable equilibria: $\overline{\pi}^{IC}_{\overline{\pi}_0} \neq \underline{\pi}^{IC}_{\underline{\pi}_0}$.

The intuition for why π^{IC} will *always* be different for different $\mathbf{S}_{\pi_0}$ is as follows: when the composite insurrection function is such that the sets of implementable coercion levels are different, it implies that some levels are sustainable and contained in one of the sets, but unimplementable from the highest implementable coercion level in the lower set. As $\rho(\cdot)$ is homogenously increasing, this implies that one of the sets must contain a higher π corresponding to a lower $q^*(\pi)$ than the other; the dynamically stable equilibrium will always be the lowest attainable $q^*(\pi)$; and π^{IC} is the minimum of $\mathbf{Q}_{\pi_0}$ and must consequently be different for the two different sets: $\overline{\mathbf{S}}_{\overline{\pi}_0}, \underline{\mathbf{S}}_{\underline{\pi}_0}$.[31]

Consequently, the model makes the prediction that coercion levels and corresponding SSE prevalence of morality groups will in some cases be inherently dependent on history in conjunction with the included long-term equilibrating factors.

The crucial assumption behind the developed results is that some levels of coercion

[31] We also showed that, equivalent to different sets of implementable coercion levels is that the lowest coercion level in the subset containing the coercion level equal to the upper dynamically stable equilibrium, $\overline{\pi}^{IC}_{\overline{\pi}_0}$, must not be implementable from the lowest implementable $q^*(\pi)$ from the other initial condition: $inf\{\mathbf{Q}_{\underline{\pi}_0}\}$. This $inf\{\mathbf{Q}_{\underline{\pi}_0}\}$ is by definition the $q^*(\pi)$ at the dynamically stable coercion level for this initial condition, $\underline{\pi}_0$. Defining the subset $\mathbf{s}'' \in \mathbf{S}_{\overline{\pi}_0}$ as the subset containing $\overline{\pi}^{IC}_{\overline{\pi}_0}$, an alternative formulation of Proposition 5 is: Dynamically stable equilibria differ, $\overline{\pi}^{IC}_{\overline{\pi}_0} \neq \underline{\pi}^{IC}_{\underline{\pi}_0}$, if and only if there exist two initial conditions $\overline{\pi}_0 \neq \underline{\pi}_0$ such that $\hat{\rho}(\underline{\pi}^{IC}_{\underline{\pi}_0}) < inf\{\mathbf{s}''\}$.

are marginally effective and that at least one level of coercion in between these levels is strictly marginally ineffective.

We note that any constraint on coercion can be categorized as either a feasibility or a strategic constraint, in the sense that it will either be endogenously dependent on $q^*(\pi)$ or it will not. Furthermore, note that any interval of coercion $[0, \pi_{max}]$ giving a class I) functional form of $q^*(\pi)$ can be seen as a feasibility constraint in some larger interval $[0, \pi_{max}]$ of class II) functional form.[32] We can sum up two main insights from the model as follows.

Main insight into coercion use: A legitimacy-maximizing authority will restrain coercion of a morality when non-convexities in the response to coercion make imposing more efficient coercion levels either strategically or feasibly unimplementable.

Main insight into path dependency: Dynamically stable equilibrium is inherently given by the history of the polity whenever there exist coercion levels that are sustainable but unimplementable from some initial conditions. [33]

Furthermore, we have established that the strategic constraints must be fix-points of the composite insurrection function, the mechanisms for how polities can converge

[32] As coercion resentment is decreasing in the last part of the interval $C''(\pi) < 0$ for $\pi \in (\hat{\pi}, \pi_{max}]$, any $q^*(\pi)$ of class I) must converge towards class II) as $\pi_{max} \to \infty$. This implies $\pi^{NC} = \pi_{max}$, as $\pi_{max} \to \infty$.

[33] Note that, whenever multiple dynamically stable equilibria exist, an exogenous temporary increase in insurrection capability can move the dynamically stable equilibrium between sets of long-term implementable coercion levels.

into state failure when coercion levels are unsustainable and the formal definition of, and some necessary conditions for, weak and strong coercion reliance.

Most results should naturally generalize for other non-linear functional forms where the marginal effectiveness of coercion varies with its level; i.e., 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 instance, assuming 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.

4. The theory applied to history

We now consider the religious homogenization in early modern Europe (1517–1685) and the Soviet secularization project (1922–1991) through the lens of the model. The dynamics of changing moralities mainly take the perspective of rational elites seeking to maximize their influence, where the sentiments of the population are seen as prerequisites to their strategies. The analysis assumes instrumental motivations rather than idealistic motivations for imposing the morality of the state upon the population. This is a simplifying assumption about intentions, which by nature will ultimately be difficult to prove or refute, as it is challenging to distinguish between how religious and ideological differences serve as motivation or justification for actions. Note that if one assumes elites to be fully intrinsically motivated, their motivation will often be to implement the authorities' preferred morality as an end in itself, which coincides fully with the proposed utility function of maximizing its

prevalence. In other words, regardless of why the authorities wish to maximize their morality, the analysis remains the same as long as the authorities see it as a policy objective to maximize the prevalence of individuals who share the morality of the state.

4.1. Religious tolerance in early modern France and the Holy Roman Empire 1517–1685

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 French kings and the Holy Roman emperors (β) built their legitimacy on the Catholic Church (b), and the introduction of Protestantism (a) posed a threat to the legitimacy of their states (Johnson and Koyama, 2013).[34] In the first part of early modernity (1517–1618), both the French and Holy Roman rulers witnessed a spread of Protestantism, combined with local nobility reforming to the Protestant faith to gain regional independence, leading to religious civil wars. This new religious cleavage enhanced existing ongoing processes of regional fights for independence and removed the possibility of polyvalent religious legitimization; in France, against the backdrop of a growing absolutist state, in the Holy Roman Empire, against the backdrop of a fragmenting empire.

The initial religious wars and periods of upheaval ended with the admission of religious rights at the Peace of Augsburg (1555) in the Holy Roman Empire and the Edict of Nantes (1598) in France. These concessions were made as the rulers real-

[34]For ease of presentation we do not distinguish between different Protestant faiths, i.e., Lutheranism and Calvinism.

ized the unproductiveness of the coercive measures ($C(\pi)$), coupled with an inability to sustain the ensuing military pressure (Wilson, 2009). As Johnson 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 was counter-productive, i.e., compatible with a micro level presence of coercion resentment.

Coercion resentment in the Holy Roman Empire

The Holy Roman Empire 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 Holy Roman Empirical authorities. Protestantism served as both a cause of and an excuse for peripheral resistance against the 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 morality of the state was seen as imperative for preserving the Empire united and under the control of the ruling Habsburg elite.

After granting Protestants (a morality) the right to practice their faith at the Peace of Augsburg, the Emperor Charles V (1516–1556) still saw 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. The Habsburgs recognized the current coercion level was in the inefficient ($\pi_{\underline{q}}, \pi^e_{\underline{q}}$) interval, but were uncertain and divided on the direction forward, and whether co-

ercion levels beyond $\pi^e_{\underline{q}}$ would trigger an insurrection.

The moderates wanted to pursue a nonconfrontational line and build legitimacy for both faiths, while the traditionalists saw the Habsburgs as having an intrinsic calling to be the champions of the Catholic faith in Europe and wished to purge the empire of Protestantism through the use of force, i.e., to increase π.[35] In the language of the model, the Habsburgs seemed to see their policy options as being either nonconfrontational or strong and confrontational with levels of coercion that they knew would be at the edge of, if not beyond, the set of implementable coercion levels $\mathbf{S}_{\pi_0}$. History would prove that the Habsburgs did not have the internal military or the external strategic position to impose the coercion levels that they attempted.

In 1618, it became clear 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, the Protestants increased their numbers (Cantoni, 2015). This is compatible with an increased investment in socialization and consequently $q^*(\pi)$, 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'

[35] Emperor Rudolf II (1572–1608) conducted a conciliatory policy towards the Protestants and saw an alliance as a way to unify the Empire (Helfferich, 2009). He was, however, an introvert and an ineffective emperor, and during his reign imperial influence deteriorated. Similarly, in his short reign Emperor Mattihas (1612–1618) ran conciliatory policies towards the Protestants.

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).

Initially, Emperor Ferdinand II was successful in suppressing the Protestants; in the terminology of the model he had overstepped his insurrection constraint, but was able to crush the ensuing rebellion. The Empire won the first part of the war ending with the Battle of White Mountain (1620). Ferdinand II might have succeeded in uniting the Empire under one faith had it not been for foreign involvement in the conflict. Foreign powers joined the conflict and sided with the Protestants to further their own causes: Protestant Denmark-Norway feared that a Catholic victory would threaten its sovereignty, Sweden feared that a strengthened emperor would ally with Catholic Poland to reclaim the Swedish crown, while the Catholic French aided the Protestant rebels in the Holy Roman Empire to weaken the Habsburg Empire and deter the elite from supporting their Habsburg relatives reigning in Spain.[36] The Holy Roman Empire lost the war; Emperor Ferdinand II's violation of the insurrection constraint of the Holy Roman Empire in 1618 is, together with the Peace of Augsburg, responsible for Germany remaining a religiously divided country to this day.

[36]Ringmar (2007) refutes this explanation of the Swedish rationale and argues that the Swedish elite embarked on the mission in Germany in order to be taken seriously as a Protestant European power. Nevertheless, the Swedes used a regional conflict to further their own agenda.

Religious cleavages worked to further other strategic agendas for neighbouring authorities, as a legitimate way to engage in foreign conflicts (Nexon, 2009). These factors were to change with the new inter-state institutional order that was established as the war ended with the Peace of Westphalia, contributing to that while Ferdinand II's attempt to remove Protestantism ultimately failed, the French Crown would succeed 40 years later.

Religious persecution of the French Huguenots

Similar to the wars of religion in the Holy Roman Empire, 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 nonconfrontational 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 nonconfrontational level of coercion, $\pi_{\underline{q}}$, in our model. The period following that of the study by Johnson and Koyama (2013) was generally a period of increased prosecution of religious minorities following increased state capacity.

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 counter-productiveness of this policy (Sutherland, 1988). Initially, 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' favours

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 1 July 1686. While it seems 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 military 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 fixed points of the composite insurrection function along the program of increases in π towards the dynamically stable equilibrium, which would turn out to be the gunpoint equilibrium. In line with Proposition 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 $(\pi_{\underline{q}}, \pi^e_{\underline{q}})$.

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). 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 points to either the "cultural memory" of persecution, $C(\pi)$, as having a long-term identity building effect, or a persistent high investment in socialization. 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 pointing to nascent expectations of minority rights being respected in international relations.

A comparative perspective on religious persecution in the Holy Roman Empire 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 be the start 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, which killed an estimated third of the population within the borders of modern-day Germany, increased both demand for a new paradigm and respect for new institutional rules of international relations in Europe.[37] Furthermore, among the elites, the Thirty Years' War was seen as an ex-

[37]Estimates of population loss range from 10 per cent to 45 per cent (Theibault, 1997).

ample 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 territory. It tied religious identities to territorial identities, increasing the need for religious homogeneity. The treaty obliged the king to have the same religious affiliation as that of his polity (Wilson, 2009), thus reducing the incentives of changing faith to gain power. Furthermore, it outlawed the use of religious tension in neighbouring 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, as religious minorities lost 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 lower prevalence of non-state morality, either through a quicker convergence towards an 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 religious schisms as a pretext for foreign involvement in internal conflicts.[38]

[38]The potentiality of foreign powers using religious schisms to legitimize military action in Continental Europe, and the absence of this risk in Britain as an island nation with strong natural borders, might potentially provide another piece of the puzzle in understanding the comparative early emergence of elite intentions to achieve religious tolerance in Britain.

While attempts to homogenize the Holy Roman Empire led to insurrection, foreign involvement and subsequent religious division of the Empire, the potential Huguenot mobilization could not be turned into a pretext for foreign involvement and consequently became a military threat to the French king under the new institutional framework. This absence of threat from neighbouring countries greatly relaxed the insurrection constraints as governments could focus on internal enemies when pursuing homogenization, thus predicting a closer alignment between territory and state moralities (Nexon, 2009).

The changing military technology, away from professionalized soldiers with training in the use of both firearms and swords, towards mass armies primarily reliant on gunpowder, placed a higher military value on draftable citizens.[39] In the language of the model, changing military technology led to a higher $\rho'(q^*(\pi))$ and the insurrection constraint became more sensitive to the prevalence of state morality as military capabilities became more sensitive to mass support. This, in turn, led to an increase in demand for homogenization of populations, enabling the drafting of large standing armies against external threats, which would propel the development of consolidated states.

4.2. The Soviet secularization project 1922–1991

The Soviet authorities (β) had a clear and stated agenda to reduce the prevalence of religious morality (a), and used coercion (π) against the major religions of the USSR,

[39]The empirical relationship between military technology and the need for mass armies is discussed widely, from the classic Roberts (1954) to the recent economics literature, see for instance Onorato, Scheve and Stasavage (2014).

Christianity and Islam, in order to increase the prevalence of its own secular morality, communism (b).[40] Similar to that of the Catholic kings of early modernity, the approach towards religious communities in the USSR was initially very oppressive. The Great Terror of the 1930s saw widespread killings and forced gulag encampment of religious individuals who failed to denounce their religion. From 1937 onwards, the Soviet authorities altered their policies towards religion. The combination of strong coercive measures which proved inefficient according to the Soviet authorities' own 1937 consensus and 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).[41]

Data from Froese (2008) shows how religious morality (q) in the USSR decreased as a consequence of deliberate Soviet policies to reduce its prevalence (see Figure 3.), while it increased again after the fall of the Soviet Union following the cessation of anti-religious policies (π) (see Figure 4.). All in all, the attempt to secularize Christian regions of Soviet society was successful in that it led to a drastic reduction in the prevalence of religious morality, but it did not lead to full removal of religious sentiment.

[40]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).

[41]Illustrative of the approach of the authorities are the names of the atheist movement founded by the USSR authorities. Before 1920 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 when the secularization project was put on hold. A subsequent atheist organization founded after WWII was named the Knowledge Society.

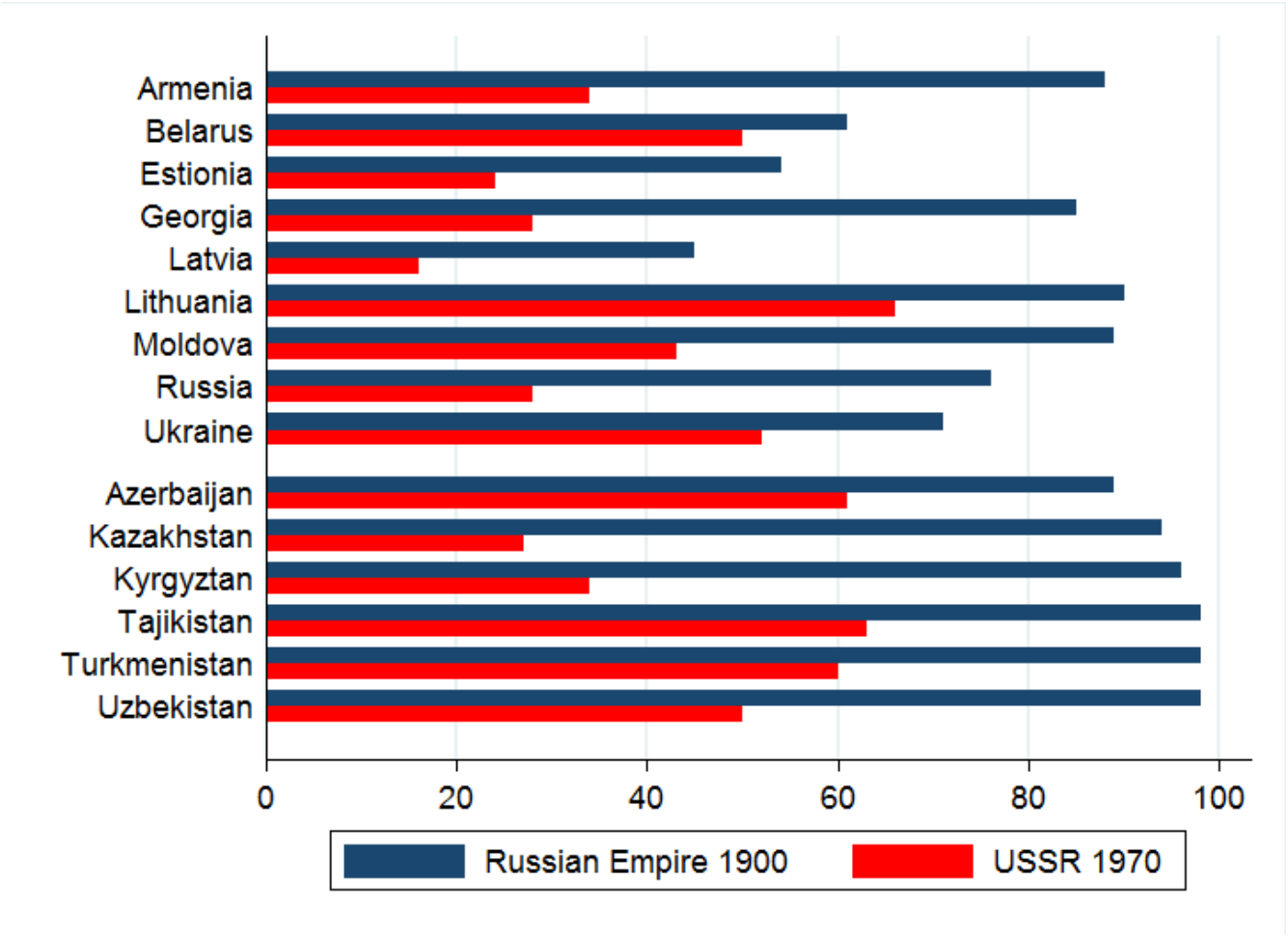


Figure 3: Religious affiliation with all religions before the Soviet Union, during the Russian Empire (1900), is shown as the bottom blue bars and in the Soviet Union (1970) is shown as the top red bars

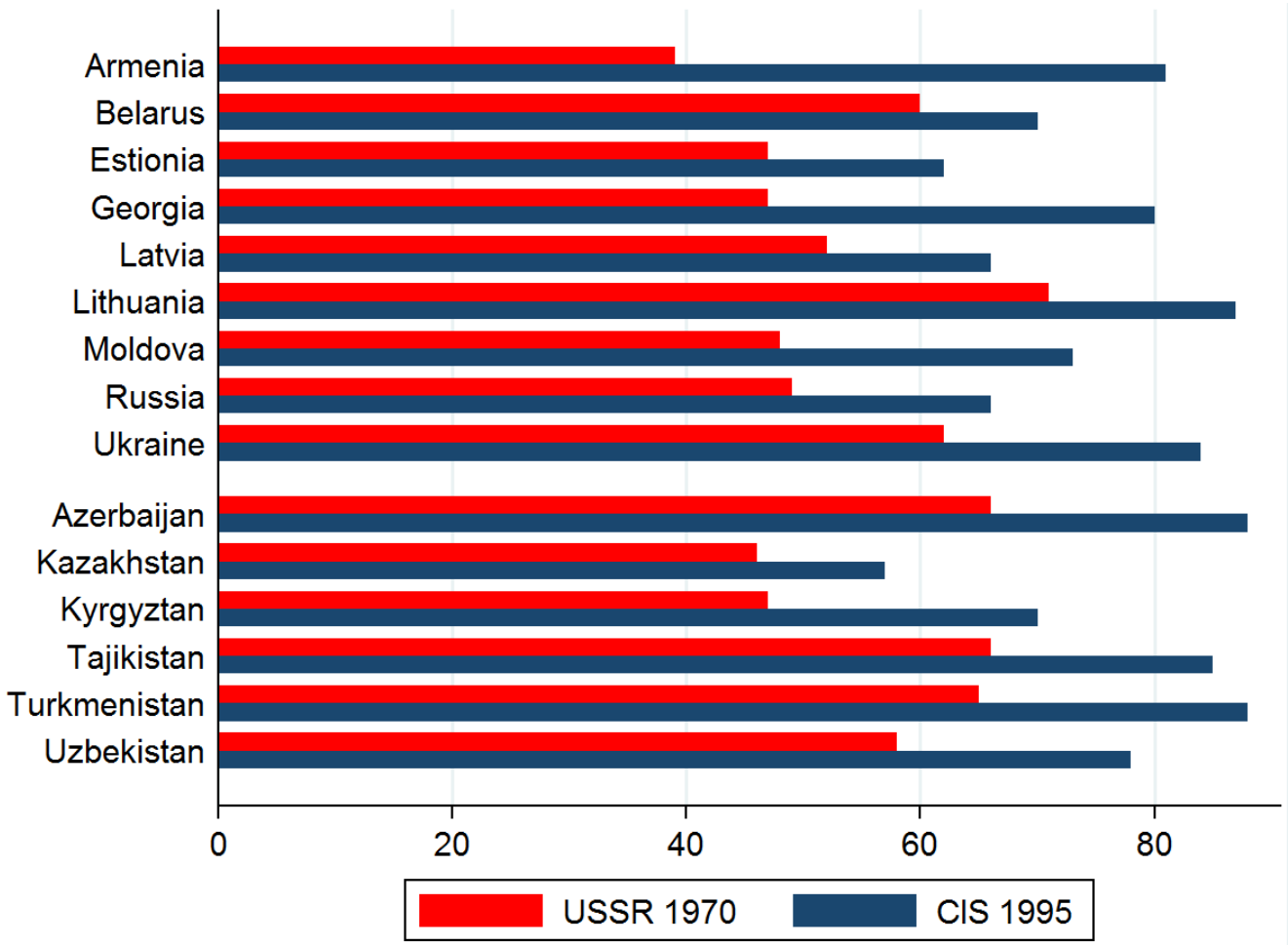


Figure 4: Affiliation with the majority religion in the Soviet Union (1970) is shown as the top red bars and after the Soviet Union in the Commonwealth of Independent States (CIS) (1995) is shown as the bottom blue bars. Source: Froese (2008). All Christian countries have Orthodox Christianity as the majority religion except Lithuania which is Roman Catholic and Latvia which is Lutheran, while all Muslim countries in the Soviet Union have Sunni Islam as their majority religion except for Azerbaijan, which is of the Shia Islamic faith.

The Soviet authorities and Christian churches

From 1937 onwards the major Christian churches of the Soviet Union were able to continue their practice, albeit facing censorship and demands from the authorities to serve the purposes of the Communist Party. The high degree of organization and internal hierarchies meant that both the Russian Orthodox Church and the Catholic Church were forced to become integrated into the Soviet system and continue their practice while facing strong censorship. Protestantism and other less hierarchical Christian communities were often strengthened by feelings of spite towards the Soviet authorities' anti-religious policies, i.e., a response to the level of coercion. To the extent that religion persisted in the predominantly Christian parts of the Soviet Union, it did so largely by the use of what Greif and Tadelis (2010) refer to as crypto-morality: hidden from the public eye.

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 especially true where the church was seen as opposing the state; one example is membership of the autonomous Lithuanian Roman Catholic Church, which was seen as synonymous with resistance to the Soviet authorities Froese (2008). This suggests that where the framing of religious persistence was aligned with other in-group and out-group dynamics, the cultural difference, Δu, towards secular USSR identity was higher. A possible explanation is that, as the framework predicts, a higher Δu leads to a higher prevalence of non-state morality, 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.

Coercion resentment in Central Asia region of the Soviet Union

The secularization policies towards the Muslims in Central Asia were even more cautious than towards the predominantly Christian Orthodox part of the Soviet Union. In Central Asia, the insurrection risk was higher because of the weaker military presence and larger cultural differences between the local Muslim morality and the secular Communist morality, Δu. The potential gains for the local population from Soviet rule, modernization and economic development, were higher than 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 to co-operate, contributing to a climate of communication between elites that was comparatively more benign than that between Moscow and Baltic elites.

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 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, Bird and Blane, 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).

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). Policies such as the removal of Muslim courts were cautiously framed as modernization and done in co-operation with moderate Islamists. Stalin initially 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 nonconfrontational, 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).

Stalin would later 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). 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 in part as a result of the anti-religious policies to gain increasing salience in 1967. They account for this effect by what they describe as "resentment against cultural and administrative domination of the Russians", sentiments that could be turned into momentum for an insurrection against the USSR.

However, the promises of development and growth were not reneged by the So-

viet authorities; they trusted that their Muslim counterparts would not attempt to secede enough to endow them with a more working economy. This growth happened alongside positive social changes in Central Asian USSR; for instance, women were given a comparatively independent role and educational levels were increased, further integrating Central Asian USSR with Moscow. Together with the strengthening of Russian military capability, these changes made any threat of cessation less realistic (Conquest, 1970).

Stalin either persecuted in a heavy-handed manner or kept a nonconfrontational approach. This implies that he avoided a mid interval of coercion, in line with the model's prediction of authorities avoiding the inefficient $(\pi_{\underline{q}}, \pi^e_{\underline{q}})$ interval where latent strategic constraints are in place. 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). The potential threat of the Russian Muslim population in the south-west, and Caucasus allying with the Germans, were used as a pretext for deportations, but this motive cannot explain why the differential treatment persisted after the end of WWII. The comparative differences may be explained by the Soviet authorities being aware of an inefficient interval of coercion $(\pi_{\underline{q}}, \pi^e_{\underline{q}})$ and restraining their use of coercion as a response to an insurrection constraint in Central Asia, while pursuing levels beyond $\pi^e_{\underline{q}}$ in Europe, where they had no strategic constraints. Assuming that Stalin perceived the response to coercion as stable across regions, this historical evidence supports Proposition 2 and indicates that the combination of cultural differences and a negative response to

coercion was sufficiently hostile as to fall under class II) in the USSR.[42]

5. Concluding remarks

We have developed a model that demonstrates how the micro foundations of coercion resentment can be used to understand the macrodynamics of legitimacy-maximizing authorities. We have argued that the assessed monarchs of Early Modernity and Stalin restrained their use of coercion in response to strategic constraints in a way that is explainable by our theoretical framework.

The model implicitly assumes atomized agents and abstracts from dynamics of legitimacy caused by communities, organizational structure, framing or strategic interaction between elites. These implicit simplifications are justified as long as community leaders are equally good at maximizing their own influence by playing on salient cleavages. If both moralities 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", i.e., the actual 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.

[42]The disparately harsher treatment of the eastern Muslims continued after the end of WWII, pointing to an additional consideration for Stalin's differential treatment, the heightened need for greater legitimacy and a stronger capability for coercion in Europe. Stalin also gained an additional benefit from coercing the eastern Muslims; it sent a credible signal about the cost of collaboration with competing authorities to other minority communities.

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 parents of state morality can fully or partially set the coercion level. In the model here developed parents of the state morality will need lower levels of parental socialization to attain their preferred morality in an equilibrium with positive levels of coercion.[43] Assuming that the majority of parents do not have utility in the outcomes of the state, parents of state morality 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 morality 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 to what extent totalitarian policies emerge from political demand or political supply.

Evolutionary properties of state competition: note how the model predictions hold in a framework where authorities are naïve about the effect of coercion; authorities that impose coercion levels within the set of implementable coercion levels will endure, and others 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. For instance, consider an extension of this model where populations of polities of uniform size and initial con-

[43]This holds as long as the state morality is also the majority morality. If the state morality is a minority, the issue depends on functional form i.e. socialization responses to the use of coercion.

ditions compete. Assume that the authorities are naïve about the effect of coercion but 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 about the effects of, and constraints on, coercion.

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 or legitimacy.[44] Furthermore, under costly coercion, 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, i.e., with more or less direct hereditariness of power, could lead to different policies. In modern

[44] 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, giving a low insurgency cost compared with the low value of attaining legitimacy, together with multiple historic influences, i.e., multiple seed moralities, and low benefits of having legitimacy, have contributed to the large cultural heterogeneity observed in Afghanistan.

democracies, such differences in incorporating the future can arise from variations between candidate politics versus party politics.

Strategic aspects of multiple competing authorities: technology and the composition of multiple ethnic groups might make the set of sustainable coercion levels empty for any single authority; hence in some instances, creating strategic coalitions between authorities is a necessity for establishing a sustainable state. The model could be extended towards a co-operative game theory framework to analyze the strategic dimensions of internal and external competing authorities under varying insurrection constraints. This can address questions such as sustainable polity borders and how intervention in polities with multiple authorities should optimally apply local power structures.[45] Furthermore, the endogenous treatment of morality prevalence enables the framework to identify a peace agreement between competing authorities that will not be sustainable in the long term; i.e., when long-term changes in the prevalence of moralities will affect power balances to render a previous agreement an out-of-equilibrium outcome.

Framing and timing of coercion: it seems 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

[45] Expert on state development in Africa, Robert H. Bates, predicts that the key to understanding the structure of wars in Africa versus Europe lies in understanding that the Peace of Westphalia constrained the European elites in terms of using neighbouring ethnic cleavages to further their cause (Weingast and Wittman, 2008). Use of ethnic conflicts in the search for influence has generally been a major cause of instability in central Africa; one example is the conflict in eastern Congo (1998–2003).

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. Explicit modelling of the effects of timing and framing of coercion can be applied to understand how short-term processes determine convergence towards the long-run equilibria.

Strategic conditions of coercion reliance: which strategic pre-histories are conducive for authorities establishing coercion reliant states? Addressing this question can complement the rich and established literature on the path to inclusive institutions from a new angle: how did authorities strongly reliant on coercion arrive in this situation? Furthermore, it can potentially give theoretical insights into which paths of state development lead to malignant outcomes, and at what critical junctures these paths can be avoided.

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 about the ratio of the cost of commission versus the cost of omission, i.e., 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, i.e., polity borders that can create sustainable uncoercive states, the cost of reaching these states and where the pitfalls of state failure lie.

Building and empirically investigating general models of these dynamics play an

important role in using the conflicts of the past to avoid conflicts in the future, and to further understand how diversity of moralities 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. From understanding democratic transitions to policy recommendations in states such as Syria and Afghanistan, legitimacy and its dynamics remain an important phenomenon.

6. Acknowledgments

First, I would like thank to my supervisors, Avner Greif and Bertil Tungodden for insightful comments, support, inspiration and help. Second, I would also like to thank Robert Arons, Thor Andreas Aursland, 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, Simen Ulsaker, Tom Grimstvedt Meling, Moti Michaeli, Linda Nøstbakken, Frederik Willumsen, participants at the 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, UIB-NHH PhD 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 encouragement. Third, I would like to thank Anne Liv Scarce, Vivienne Bowery Knowles and Karin Lillevold for research assistance. I gratefully acknowledge support from the Research Council of Norway(NFR) through the FAIR Centre, NFR grant number 262675.

7. Appendix 1: Proofs

7.1. Proof of Lemma 1

Lemma 1: The smallest group always invests more in parental socialization: $\tau_t^b \leq \tau_t^a$ if and only if $q_t \leq (1 - q_t)$.

Proof: Suppose $q_t < (1 - q_t)$, it then follows from (6) that $H'(\tau_t^b) < H'(\tau_t^a)$. By the Inada condition of $H''(\tau^m) \geq 0$ in Assumption 3, it follows that $\tau_t^b < \tau_t^a$. The only if part follows from the fact that there are only two moralities.
□

7.2. Proof of Lemma 2

Lemma 2: 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 (3), $q_{t+1} = q_t + q_t(1 - q_t)(\tau^a - \tau^b)$ that for $q_t = q_{t+1}$ to be fulfilled, $q_t(1 - q_t)(\tau^a - \tau^b) = 0$ must hold. Hence, at any interior SSE, i.e., $q^* \in (0, 1)$, $\tau^a = \tau^b$. From (6), 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 (6) that $H'(\tau_t^b) < H'(\tau_t^a)$. By Lemma 1 it follows that $\tau_t^b > \tau_t^a$. By (3), $q_{t+1} < q_t$ when $\tau_t^b < \tau_t^a$ and $q_{t+1} < q_t$ for $\tau_t^b > \tau_t^a$. Thus,

any $q_t \in (0,1)$ will converge to q^*. In other words, $(0,1)$ is a q^* basin of attraction. □

7.3. Proof of Lemma 3

Lemma 3: 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 $q^*(\pi) = \frac{\Delta u - \pi + C(\pi)}{2\Delta u + C(\pi)}$.

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 $q_t \in \{0,1\}$, (3) 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 $\Delta u' - \pi' + C(\pi') = 1 > 0$, (10) implies that $\tau_a > 0$ for $q_t > 0$. As $\Delta u' + \pi' > 0$, we see from (11) that $\tau_b > 0$ for $(1 - q_t) > 0$. For $q_t = \frac{1}{2}$ to be an SSE, it follows from (3) that $\tau_a = \tau_b$. This implies the left side of (10) should equal the left side of (11). 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 $\frac{\Delta u' - \pi' + C(\pi')}{2\Delta u' + C(\pi')} = \frac{1}{2}$.[46]

Uniqueness of the interior SSE $q^*(\pi)$ is trivially given by the fact that $q^*(\pi) =$

[46]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.

$\frac{\Delta u-\pi+C(\pi)}{2\Delta u+C(\pi)}$ 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 2 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 (10) 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 (11) that $\tau_b > 0$ for all q_t. It follows from (3) that if $\tau_a \neq \tau_b$ for all q_t, no interior SSE exists.

□

7.4. Proof of Lemma 4

Lemma 4: 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: 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 Lemma 3, 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 $H'(\tau_t^b) \neq H'(\tau_t^a)$; because of the Inada conditions on $H(\cdot)$ it follows that $\tau_t^b \neq \tau_t^a$, and by (3) it follows that $q_1 \neq q^*(\pi_0)$. We define the following sequence of $q_0, q_1 ... q_N$ values under π' as $\{q_0, q_1 ..... q_N, \pi'\} \equiv \{q_t\}_{\pi'}$.

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 proof of Lemma 3, it holds that $(1-q^*(\pi'))(\Delta u - \pi' + C(\pi')) = q^*(\pi')(\Delta u + \pi')$, as follows from the proof of Lemma 3. Suppose $q_t > q^*(\pi')$. It then follows that $(1-q_t)(\Delta u - \pi' + C(\pi')) < q_t(\Delta u + \pi')$ which, by (10) and (11), implies $H'(\tau_a) < H'(\tau_b)$. It follows from the Inada condition of $H'(\cdot) > 0$ that this implies $\tau_t^b > \tau_t^a$. Suppose $q_t < q^*(\pi')$, then the opposite holds. By (3) it holds that $q_t < q_{t+1}$ when $\tau_t^b < \tau_t^a$ and $q_t > q_{t+1}$, if $\tau_t^b > \tau_t^a$.

(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 $\{q_t\}_{\pi'}$.

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, (10) and (11) and the Inada condition $H'(0) = 0$, implies 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 $\tau_t^a > 0, \tau_t^b > 0$ 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 (3), $q_{t+1} = q_t + q_t(1-q_t)(\tau^a - \tau^b)$. We see that an exterior $q^* = 0$ or $q^* = 1$ cannot be reached from any interior $q_t \in (0,1)$ if $\tau^a > 0, \tau^b > 0$.

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

This proof applies Proposition 1, Proposition 2 and the definition of cultural substitution in Bisin and Verdier (2001 p 303–307). Following the proof of Proposition 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 Proposition 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 morality 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 (10) and (11) it follows that:

$$\tau_a = H'^{-1}((1 - q_t)(\Delta u - \pi' + C(\pi')). \tag{19}$$

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

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 (19), (20) but q_t remains fixed for all $q_t \in \{q_t\}_{\pi'}$, and because $q^*(\pi') > 0$, implies that $\Delta u - \pi' + C(\pi') = K^a > 0$, $\Delta u + \pi' = K^b > 0$;

we can define $H'^{-1}(q^m K^m) \equiv d^m(q^m)$. As $q_t \in (0,1)$, we can write $\tau^a = d^a(q_t)$ and $\tau^b = d^b(1-q_t)$.

(ii) $d^m(1) = 0$.
Following from (8) and (9), 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 $\tau^a = 0$ for $q_t = 1$ and $\tau^b = 0$ 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 (19) and (20) 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 we have established that $H'^{-1}(\cdot)$ is continuously defined, therefore 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 Proposition 1. Inserting $\tau^a = d^i(q_t)$ and $\tau^b = d^i(1-q_t)$ into (3) 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:[47]

$$\dot{q} = q(1-q)[d^a(q) - d^b(1-q)]. \tag{21}$$

From part *(i)* of the proof we have $\tau^a > \tau^b$ 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

[47] 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.

$\frac{\partial \dot{q}_t}{\partial q}\Big|_{q=0} = d^a(0) - d^b(1) > 0$ 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 (21) continuously maps from q into $\dot{q}$, the basin of attraction of $q^*(\pi')$ under π' is $(0, 1)$, which implies that $q_t \to q^*(\pi')$ (Bisin and Verdier, 2001).

□

7.5. Proof of Lemma 5

Lemma 5: Coercion is marginally effective at the beginning and end of $[0, \pi_{max}]$, and there is at least one level of coercion, $\hat{\pi}$, that is strictly marginally ineffective: $q^{*\prime}(\hat{\pi}) > 0$.

Proof: Marginal effectiveness is defined as $q^{*\prime}(\pi) < 0$. We show that Assumption 5 implies marginal effectiveness is negative at $\hat{\pi}$, and that marginal effectiveness of coercion is positive at $\pi \in \{0, \pi_{max}\}$. We first show that $q^{*\prime}(\pi) < 0$ for $\pi \in \{0, \pi_{max}\}$. Generally, $q^{*\prime}(\pi)$ is given by:

$$q^{*\prime}(\pi) = \frac{(C'(\pi) - 2)\Delta u - C(\pi) + \pi C'(\pi)}{(C(\pi) + 2\Delta u)^2}. \tag{22}$$

Assumption 6 implies $C'(\pi) = 0$ at $\pi \in \{0, \pi_{max}\}$. Inserting $C'(\pi) = 0$ into (19) gives:

$$q^{*\prime}(\pi) = \frac{-1}{C(\pi) + 2\Delta u} < 0. \tag{23}$$

We now show that $q^{*\prime}(\hat{\pi}) > 0$. By the functional form assumption on (13) $C''(\hat{\pi}) = 0$ and by the C^2 assumption on $C(\pi)$ in (12), there must be an open interval of ϵ length around $\hat{\pi}$, where $C''(\pi) = 0$. In this interval the function is linear, hence we can assume the coercion resentment function is $C(\pi) = K_0 + K_1\pi$ for some K_1. For

$C(\pi) = K_0 + K_1\pi$ the marginal effectiveness of coercion in π', $q^{*\prime}(\pi')$, is given by:

$$q^{*\prime}(\pi') = \frac{(K_1 - 2)\Delta u}{(K_1\pi' + 2\Delta u)^2}. \tag{24}$$

As the functional form assumptions of $C(\pi)$ made in (13) imply that $C''(\pi)$ is strictly non-zero in $[0, \hat{\pi})$ and negative in $(\hat{\pi}, \pi_{max}]$, it must be that $C'(\pi)$ is at its maximum value at $\hat{\pi}$. By the last part of Assumption 6 at least one $\pi' \in (0, \pi_{max})$ must have $C'(\pi) > 2$, and the maximum of $C'(\pi)$ is at $\hat{\pi}$. Hence, it must hold that $C(\hat{\pi}) > 2$. Inserting $K_1 = 2$ into (20), we attain $q^{*\prime}(\pi) = 0$, and because $K_1 = C(\hat{\pi}) > 2$ it must hold that $q^{*\prime}(\hat{\pi}) > 0$.

□

7.6. Proof of Lemma 6

Lemma 6: $q^*(\pi)$ 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 $q^{*\prime}(\pi)$ over $[0, \pi_{max}]$ must be; we then use this to show (iii) that any possible $q^*(0) + \int_0^{\pi_{max}} q^{*\prime}(\pi)d\pi$ 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 $q^{*\prime}(\pi) = 0$ in $[0, \hat{\pi})$. We note from the proof of Lemma 5 that:

$$q^{*\prime}(\pi) = \frac{(C'(\pi) - 2)\Delta u + \pi C'(\pi) - C(\pi)}{(C(\pi) + 2\Delta u)^2}. \tag{25}$$

It follows from Lemma 5 that $q^{*\prime}(0) < 0$, which by (25) implies $(C'(0) - 2)\Delta u < C(0)$. Similarly, it follows from Lemma 5 that $q^{*\prime}(\hat{\pi}) > 0$ which by (25) 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 (12), hence $(C'(\pi) - 2)\Delta u + \pi C'(\pi)$ and $C(\pi)$ must cross at $(0, \hat{\pi})$, giving $q^{*\prime}(\pi) = 0$ 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, (12) and (13), 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 $q^{*\prime}(\pi_{\underline{q}}) = 0$ 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 (25), we see that $q^{*\prime}(0) > 0$, $q^{*\prime}(\hat{\pi}) < 0$ 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 (25), we see this implies $q^{*\prime\prime}(\pi) > 0$ for $\pi \in (\pi_{\underline{q}}, \hat{\pi}]$, $q^{*\prime}(\pi) > 0$ for any $\pi > \pi_{\underline{q}}$, and $q^{*\prime}(\pi) < 0$ 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 (12) and (13) 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 Lemma 5 it follows that $q^{*\prime}(\hat{\pi}) > 0$ and $q^{*\prime}(\pi_{max}) < 0$. Hence, $q^{*\prime}(\pi)$ 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 $q^{*\prime}(\pi') = 0$. This π' is defined as $\pi_{\overline{q}}$, the unique maximum in $(\hat{\pi}, \pi_{max}]$.[48]

(ii) The sign of $q^{*\prime}(\pi)$

First, note that from part (i) of the proof we have $\pi_{\underline{q}} < \hat{\pi} < \pi_{\overline{q}}$. From Lemma 5 and part (i) of the proof, it follows that $q^{*\prime}(\pi)$ is strictly increasing from $q^{*\prime}(0) < 0$ to $q^{*\prime}(\pi_{\underline{q}}) = 0$ and onward to $q^{*\prime}(\hat{\pi}) > 2$, and strictly decreasing from $q^{*\prime}(\hat{\pi}) > 2$ to $q^{*\prime}(\pi_{\overline{q}}) = 0$ and onward to $q^{*}(\pi_{max}) < 0$. Hence, if $q^{*}(\pi) > 0$ for all $\pi \in [0, \pi_{max}]$, then we note the following.

$$q^{*\prime}(\pi) < 0 \quad \text{for all} \quad \pi \in [0, \pi_{\underline{q}})$$

$$q^{*\prime}(\pi) > 0 \quad \text{for all} \quad \pi \in (\pi_{\underline{q}}, \pi_{\overline{q}})$$

$$q^{*\prime}(\pi) < 0 \quad \text{for all} \quad \pi \in (\pi_{\overline{q}}, \pi_{max}]$$

[48] 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 $q^{*}(\pi_{\overline{q}}) \in (0, 1)$.

(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 $q^{*\prime}(0) < 0$ and $q^{*\prime\prime}(0) > 0$. If $q^*(\pi'') = 0$ for some π'' in the interval $[0, \hat{\pi})$, where $q^{*\prime}(\pi) < 0$ such that $q^*(0) + \int_0^{\pi''} q^{*\prime}(\pi)d\pi = 0$ then, because $q^{*\prime}(\pi_{\underline{q}}) = 0$ and $q^{*\prime}(\pi_{\underline{q}}) > 0$ 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 $q^{*\prime}(\pi) < 0$ such that $q^*(\pi'') = 0$ exists, then $q^{*\prime}(\pi') = 0$ where $q^*(\pi') > 0$ exists, and this π' is $\pi_{\underline{q}}$. As $q^{*\prime}(\pi) > 0$ 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 $q^{*\prime}(\pi) = 0$ for any $q^*(\pi) = 0$ such that we can define integrals of $q^{*\prime}(\pi)$ 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.

$$q^*(0) + \int_0^{\pi_{\underline{q}}} q^{*\prime}(\pi)d\pi + \int_{\pi_{\underline{q}}}^{\pi_{\overline{q}}} q^{*\prime}(\pi)d\pi + \int_{\pi_{\overline{q}}}^{\pi_{max}} q^{*\prime}(\pi)d\pi \tag{26}$$

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

We note the definition of $\pi^e_{\underline{q}}$ is $\pi_{\underline{q}} < \pi^e_{\underline{q}}$ and $q^*(\pi^e_{\underline{q}}) \equiv q^*(\pi_{\underline{q}})$. If $\int_{\pi_{\underline{q}}}^{\pi_{\overline{q}}} q^{*\prime}(\pi)d\pi +$

$\int_{\pi_{\overline{q}}}^{\pi_{max}} q^{*\prime}(\pi)d\pi \leq 0$, then, because all functions are continuous, $q^*(\pi^e_{\underline{q}})$ must be defined. Considering (26) and part (ii) of the proof, we see this implies that $q^*(\pi)$ has two minima, $\pi_{\underline{q}}$ and some $\pi' : \pi' \geq \pi^e_{\underline{q}}$, and one interior maximum, $\pi_{\overline{q}}$. This implies that the functional form falls within class II).

If $\int_{\pi_{\underline{q}}}^{\pi_{\overline{q}}} q^{*\prime}(\pi)d\pi + \int_{\pi_{\overline{q}}}^{\pi_{max}} q^{*\prime}(\pi)d\pi > 0$, then there will be no $q^*(\pi')$ where $\pi' > \pi_{\underline{q}}$; i.e., $\pi^e_{\underline{q}}$ is not defined. Considering (26), 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 $q^{*\prime}(\pi)$ 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 $q^{*\prime}(\pi)$ 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 $\int_{\pi_{\underline{q}}}^{\pi_{\overline{q}}} q^{*\prime}(\pi)d\pi + \int_{\pi_{\overline{q}}}^{\pi_{max}} q^{*\prime}(\pi)d\pi > 0$. 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 $\int_{\pi_{\underline{q}}}^{\pi_{\overline{q}}} q^{*\prime}(\pi)d\pi + \int_{\pi_{\overline{q}}}^{\pi_{max}} q^{*\prime}(\pi)d\pi = 0$, then $\pi_{\underline{q}}$ is a non-unique global minimum, because it must then hold that $q^*(\pi_{\underline{q}}) = q^*(\pi_{max})$. Suppose $\int_{\pi_{\underline{q}}}^{\pi_{\overline{q}}} q^{*\prime}(\pi)d\pi + \int_{\pi_{\overline{q}}}^{\pi_{max}} q^{*\prime}(\pi)d\pi < 0$, 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 $\pi' \in [\pi^e_{\overline{q}}, \pi_{max}]$ is a global minimum. As $q^{*\prime}(\pi) < 0$ 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 $\pi_{\underline{q}}$.

□

7.7. Proof of Proposition 1

Proposition 1: Let π' denote a level of coercion such that $q^*(\pi') = 0$. The optimal level of coercion under no constraint, π^{NC}, will be as follows for the different classes of $q^*(\pi)$:

I) $\pi^{NC} = \pi_{\underline{q}}$,

II) $\pi^{NC} = \pi' \in (\pi^e_{\underline{q}}, \pi_{max})$, if no π' is defined then $\pi^{NC} = \pi_{max}$,

III) $\pi^{NC} = \pi' < \pi_{\underline{q}}$, where π' is always defined.

Proof: This proof follows from Lemma 6 and the assumption that the authority is minimizing $q^*(\pi)$, as captured in (7).

□

7.8. Proof of Corollary of Proposition 1

Corollary of Proposition 1: An authority facing no constraint on coercion will only restrain its use of coercion when SSE $q^*(\pi)$ is of class I).

Proof: A restraint on coercion implies a coercion level π' being strictly smaller than the highest implementable coercion level with a non-zero $q^*(\pi)$. It follows directly from Proposition 1 that this only occurs under $q^*(\pi)$ of class I).
□

7.9. Proof of Proposition 2

Proposition 2: 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 proposition.

If the authority is minimizing $q^*(\pi)$ by (7) and Lemma 6, 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 6, it holds that $q^{*\prime}(\pi) < 0$ 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 6 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 (26) and the proof of Lemma 6 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 6, $q^{*\prime}(\pi) < 0$ 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 6 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 (14), 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.

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

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

Note that the proposition states that $\rho \neq \pi^e_{\underline{q}}$, and I) implies that $\rho = \pi_{\underline{q}} \rightarrow \pi^{EC} = \pi_{\underline{q}} = \rho$. Thus, (28) and (27) cover all possible scenarios of π^{EC} for $\rho \in [0, \pi_{max}]$, which implies the following.

$$\pi^{EC} = \rho \rightarrow \rho \notin (\pi_{\underline{q}}, \pi^e_{\underline{q}}) \tag{29}$$

$$\pi^{EC} = \pi_{\underline{q}} < \rho \rightarrow \rho \in (\pi_{\underline{q}}, \pi^e_{\underline{q}}) \tag{30}$$

Part (i) of the proposition follows from (27) and (29). Part (ii) of the proposition follows from (28) and (30).

□

7.10. Proof of Corollary of Proposition 2

Corollary of Proposition 2: An authority will restrain its use of coercion as a response to a constraint if and only if the $q^*(\pi)$ is of class II) and the constraint is in the inefficient interval of coercion, $\rho \in (\pi_{\underline{q}}, \pi^e_{\underline{q}})$.

Proof: We first note that by definition on page 28, an authority exhibiting restraint as a response to the constraint, imposes π^{EC}, that is:

1. a response to a constraint: a π^{EC} that is different than the optimal adjustment without constraints, $\pi^{EC} \neq \pi^{NC}$

2. a restraint: a π^{EC} strictly lower than its highest implementable level, $\pi^{EC} < \rho \leq \pi_{max}$.

We first show the if part, that $q^*(\pi)$ is of class II) and $\rho \in (\pi_{\underline{q}}, \pi^e_{\underline{q}})$ implies a restraint 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 restraint 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 Proposition 2 $\pi^{EC} = \pi_{\underline{q}} < \rho$. Then π^{EC} is a restraint as a response to a constraint, because $\pi^{EC} = \pi_{\underline{q}} < \pi^e_{\underline{q}} < \pi^{NC}$.

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 Proposition 2; hence, π^{EC} is not a restraint. 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 from propositions 1 and 2 we have $\pi^{NC} = \pi^{EC} = \pi_{\underline{q}}$. 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 by propositions 1 and 2 $\pi^{NC} = \pi^{EC} = \pi' < \rho$ where $q^*(\pi') = 0$; consequently, π^{EC} is not a response to a constraint.
□

7.11. Proof of Proposition 3

Proposition 3: Weak coercion reliance occurring without strong coercion reliance implies a strictly marginally inefficient interval of coercion between intervals of marginal effective coercion levels.

Proof: Strong coercion reliance implies $\pi = 0$ is an unsustainable level of coercion. Weak coercion reliance implies there exists $\pi' < \pi_0$ such that $\pi' \in [0, \pi_0)$ is unsustainable. Weak coercion reliance without strong coercion reliance implies $\pi = 0$ is a sustainable level of coercion, $0 \geq \rho(q^*(0))$, while there exists some level $0 < \pi' < \pi_0$ that is unsustainable; i.e., $\rho(q^*(\pi')) < \pi'$. As π_0 by definition implies an interior SSE not breaching the insurrection constraint, and by Assumption 7 $\rho'(\cdot) < 0$, weak coercion reliance occurring without strong coercion reliance must imply there exists a π' such that $\rho(q^*(\pi')) < \rho(q^*(0)) \leq \rho(q^*(\pi_0))$. As $\rho'(\cdot) < 0$, we have $q^*(\pi') > q^*(\pi_0)$, where $\pi' < \pi_0$.

$\pi' < \pi_0$ while $q^*(\pi') > q^*(\pi_0) \geq q^*(0)$ cannot occur without an interval of π, such

that $q^{*\prime}(\pi) < 0$ in between intervals of π such that $q^{*\prime}(\pi) > 0$, which is from the definition of marginal efficiency on page 21, is a marginal inefficient interval preceded and followed by marginal efficient levels of coercion.

□

7.12. Proof of Proposition 4

Proposition 4: For any initial condition π_0, the dynamically stable equilibrium π^{IC} is a coercion level equal to either:

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

Proof: We show that the π' corresponding to any minimum point of any $\mathbf{Q}_{\pi_0}$, which by definition is equal to π^{IC}, will fall under either case I), II) or III), hence the proposition.[49] First note that trivially, any minimum point in $\mathbf{Q}_{\pi_0}$ must correspond to a π^{IC} in the interior of a subset of an $\mathbf{S}_{\pi_0}$, $\mathbf{s}$, or at the boundary of an $\mathbf{s}$.

Suppose the minimum of $\mathbf{Q}_{\pi_0}$ corresponds to an interior point in an $\mathbf{s}$. As established in Lemma 6, $q^*(\pi)$ has at most one interior minimum point, $\pi_{\underline{q}}$, and because $\rho'(q^*(\pi)) < 0$ always holds, $q^{*\prime}(\pi) = 0$ must hold at a minimum of $\mathbf{Q}_{\pi_0}$, corresponding to an interior minimum of $\mathbf{s}$. Thus, π^{IC} must be equal to $\pi_{\underline{q}}$ and π^{IC} fall under case I).

Suppose the minimum of $\mathbf{Q}_{\pi_0}$ corresponds to a π^{IC} that is the limit of a subset

[49] If there are several infimum points, any will correspond to a dynamically stable equilibrium, as the authority will not have any incentive to change π.

$\mathbf{s}$, and this limit is different from π_{max}. π^{IC} must be at an upper limit of $\mathbf{s}$, because at lower thresholds of $\mathbf{s}$ lowering π increases $q^*(\pi)$, as follows from proof of Proposition 3. As the limit π^{IC} is an upper limit different from π_{max}, it implies there exists a $\pi^{IC} < \pi' < \pi_{max}$ such that π' is infinitesimally larger than the upper limit of the subset. As $\pi' \notin \mathbf{s}$, $\rho(q^*(\pi')) > \pi'$. As $\rho(\cdot)$ is assumed to be a continuous mapping with a continuous derivative, it cannot discontinuously jump from π^{IC}, 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.[50] Hence, the minimum of $\mathbf{Q}_{\pi_0}$ must correspond to an upper limit on the fix-point-line $\pi^{IC} = \overline{\pi_{fix}}$, which falls under case II).

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

7.13. Proof of Proposition 5

Proposition 5: 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, $\overline{\mathbf{S}}_{\overline{\pi}_0} \triangle \underline{\mathbf{S}}_{\underline{\pi}_0} \neq \emptyset$ will different initial conditions $\pi_0 = \overline{\pi}_0$ and $\pi_0 = \underline{\pi}_0$ give different dynamically stable equilibria: $\overline{\pi}^{IC}_{\overline{\pi}_0} \neq \underline{\pi}^{IC}_{\underline{\pi}_0}$.

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

Suppose $\overline{\pi}_0 > \underline{\pi}_0$ and $\overline{\mathbf{S}}_{\overline{\pi}_0} \triangle \underline{\mathbf{S}}_{\underline{\pi}_0} \neq \emptyset$. By definition of the set of implementable

[50] The fix-point-line for $\hat{\rho}(\pi)$ is illustrated in Figure. 2 on page 34.

coercion levels there must be at least one π' such that $\pi' \notin \underline{\mathbf{S}}_{\underline{\pi}_0}$, but $\pi' \in \overline{\mathbf{S}}_{\overline{\pi}_0}$ because if this was not the case, then the sets would be the same sets; i.e., this is implied by $\overline{\mathbf{S}}_{\overline{\pi}_0} \triangle \underline{\mathbf{S}}_{\underline{\pi}_0} \neq \emptyset$. 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 $\underline{\pi}_0$. Suppose that the difference between the sets consists of this single coercion level π'. This $\pi' > sup\{\mathbf{S}_{\underline{\pi}_0}\}$ must then be equal to $\overline{\pi}^{IC}$, because π' must corresponds to $inf\{\mathbf{Q}_{\overline{\pi}_0}\}$. This π' is different from $\underline{\pi}^{IC}$ because π' is not in $\mathbf{S}_{\underline{\pi}_0}$.[51]

Suppose the dynamically stable equilibria are different and that $\overline{\pi}^{IC}_{\overline{\pi}_0} > \underline{\pi}^{IC}_{\underline{\pi}_0}$.[52] By definition $\overline{\pi}^{IC}_{\overline{\pi}_0}$ corresponds to $inf\{\mathbf{Q}_{\overline{\pi}_0}\}$. As $\rho'(\cdot) < 0$, it must hold that $q^*(\overline{\pi}^{IC}_{\overline{\pi}_0}) < q^*(\underline{\pi}^{IC}_{\underline{\pi}_0})$ because $\underline{\pi}^{IC}_{\underline{\pi}_0}$ can be implemented from $\overline{\pi}^{IC}_{\overline{\pi}_0}$, but $\overline{\pi}^{IC}_{\overline{\pi}_0}$ gives a lower $q^*(\pi)$ than $\underline{\pi}^{IC}_{\underline{\pi}_0}$, by definition. Hence, there must be at least one π' such that $\pi' \notin \underline{\mathbf{S}}_{\underline{\pi}_0}$ but $\pi' \in \overline{\mathbf{S}}_{\overline{\pi}_0}$, namely $\overline{\pi}^{IC}_{\overline{\pi}_0}$. By definition of the set of implementable coercion levels, this implies $\overline{\mathbf{S}}_{\overline{\pi}_0} \triangle \underline{\mathbf{S}}_{\underline{\pi}_0} \neq \emptyset$.

□

8. Appendix 2: Linear coercion resentment functions

Assuming the coercion resentment function is linear, $C(\pi) = K_0 + K_1\pi$ for some K_0, K_1, yields:

$$\frac{\partial q^*(\pi)}{\partial \pi} = \frac{(K_1 - 2)\Delta u}{(K_1\pi + 2\Delta u)^2}. \tag{31}$$

[51] Note that in the special case where there are two dynamically stable equilibria but equal sets of implementable coercion levels, it is more *likely* that different equilibria will arise when there are different initial conditions.

[52] Note that the proposition does not cover the special case of multiple dynamically stable equilibria, $\overline{\pi}^{IC}_{\overline{\pi}_0} = \underline{\pi}^{IC}_{\overline{\pi}_0} = \{\pi_{\overline{q}}, \pi^e_{\overline{q}}\}$.

Using coercion as a means to change $q^*(\pi)$ would simply not be a useful tool for $K_1 = 2$, counter-productive for $K_1 > 2$, or productive at any level for $K_1 < 2$. The solution to the authority's minimization problem in (7) is trivial for linear coercion functions: either always coerce as hard as possible or never coerce at all, depending on whether the coercion resentment, K_1, is greater or smaller than two. Similarly, 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:

$$\Delta u = \frac{2C(\pi') - \pi'}{2 + 3C'(\pi')}. \tag{32}$$

9. Appendix 3: Further interpretations of the model

9.1. Appendix 3.1: Policy implications of Proposition 2

Proposition 2 has relevant implications for the policy problem of an external agency setting a constraint ρ to limit an authority's use of coercion when $q^*(\pi)$ is of class II).

Assume that the cost of enforcing the constraint is $K_0(\pi_{max} - \rho)$, where $K_0 > 0$. Setting a $\tilde{\rho} \in (\pi_{\underline{q}}, \pi^e_{\underline{q}})$ will have several benefits relative to the constraint $\rho \notin (\pi_{\underline{q}}, \pi^e_{\underline{q}})$. Imposing $\tilde{\rho}$ implies a costless reduction of the equilibrium coercion level, because the cost of imposing $\tilde{\rho}$ is $K_0(\pi_{max} - \tilde{\rho})$, and Proposition 2 implies that at $\tilde{\rho}$ the imposed coercion level π^{EC} is $\pi_{\underline{q}}$ and the actual reduction of coercion is $(\pi_{max} - \pi_{\underline{q}})$. Hence, the cost of reduction is given by $(\pi_{max} - \tilde{\rho})$ while the actual reduction is $(\pi_{max} - \pi_{\underline{q}})$, implying that the reduction $(\pi_{max} - \pi_{\underline{q}}) - (\pi_{max} - \tilde{\rho}) = (\tilde{\rho} - \pi_{\underline{q}})$ is achieved at no cost. Furthermore, assume that the external agency has imperfect information about the coercion level, giving the authority a possibility of increasing π without the external agency being able to identify the increase. At $\pi_{\underline{q}}$ the authority has no incentive to marginally increase π in equilibrium, as this would imply imposing a coercion level

π in the inefficient interval, i.e., $\pi \in (\pi_{\underline{q}}, \pi^e_{\underline{q}})$. A final benefit is that the $\tilde{\rho}$ constraint is not binding in equilibrium, which will often reduce its salience.

9.2. Appendix 3.2: 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 morality always increases their capability for committing a successful insurrection. The lower threshold for committing an insurrection 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 a morality 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.

Applying the model to a democratic system, the endogenous constraint will reflect a situation where an authority faces an undesirable outcome contingent on the level of $q^*(\pi)$ not moving beyond some threshold needed to issue a forced referendum or a motion of no confidence.

10. Appendix 4: The set of implementable coercion levels

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 (4.1), and iterative processes for $\mathbf{S}_{\pi_0}$ where π can be set at any t, in Appendix (4.2).

10.1. Appendix 4.1: 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 4, we denote this sequence of q_t values as $\{q_t\}_{\pi'}$. Now we discuss when the following criterion is met:

$$\text{If } \rho(q^*(\pi_0)) \geq \pi' \text{ and } \rho(q^*(\pi')) \geq \pi' \text{ then } \rho(q_t) \geq \pi' \text{ for all } q_t \in \{q_t\}_{\pi'}. \tag{33}$$

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 (3) 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:

$$\Delta_t \tau^m \geq \Delta_t \tau^m \Delta_{t+1} \tau^m + \Delta_{t+1} \tau^m \rightarrow 1 \geq \Delta_{t+1} \tau^m [1 + \frac{1}{\Delta_t \tau^m}]. \tag{34}$$

If (34) holds for all possible combinations of moving from one $q_t \in (0, 1)$ to another $q_{t+1} \in (0, 1)$, then (33) 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 $\overline{q}_{N,t} \equiv \frac{\Sigma_{i=0}^{N} q_{t-i}}{N+1}$. Furthermore, assume that the convergence process from $q^*(\pi_0)$ to $q^*(\pi')$ in (33) occurs within T periods. We know that $q^*(\pi')$ and $q^*(\pi_0)$ is sustainable, and it follows from proof of Lemma 4 part (iii) that any average will converge towards $q^*(\pi')$; i.e., it holds that $\overline{q}_{N,T} \rightarrow q^*(\pi')$ as $N \rightarrow \infty$. Hence, in the model as specified, (33) holds for an infinite moving average, i.e., $N = \infty$, infinite inertia. More simply put, (33) 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, (34) 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 (34). 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, $T = 1$, (33) 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. Furthermore, we have argued that it appears that the set of implementable coercion levels for an insurrection constraint dependent on $\overline{q}_{N,t}$, $\tilde{\mathbf{S}}_{\pi_0,N,T}$ will converge towards $\mathbf{S}_{\pi_0}$, 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, $\frac{1}{T} \to 1$.

10.2. Appendix 4.2: 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 4.1, 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 $\mathbf{S}_{\pi_0}$, 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, $\mathbf{S}_{\Pi}$, 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 $\overline{\pi_{fix}}$. In other words, $\mathbf{S}_{\pi_0}$ might be different for other iterative processes, but it appears that all established results would hold qualitatively.

11. References

Acemoglu, Daron and Alexander Wolitzky. 2014. "Cycles of conflict: An economic model." *The American Economic Review* 104(4):1350–1367.

Acemoglu, Daron and James A. Robinson. 2017. The Emergence of Weak, Despotic and Inclusive States. Working Paper 23657 National Bureau of Economic Research.

Alesina, Alberto and Bryony Reich. 2013. Nation Building. Technical report National Bureau of Economic Research.

Alesina, Alberto and Enrico Spolaore. 2003. *The Size of Nations*. MIT Press.

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

Barfield, Thomas J. 2010. *Afghanistan: A Cultural and Political History*. Princeton University Press.

Bennigsen, Alexandre and Chantal Lemercier-Quelquejay. 1967. *Islam in the Soviet Union*. Pall Mall Press.

Bisin, Alberto, Eleonora Patacchini, Thierry Verdier and Yves Zenou. 2011. "Formation and persistence of oppositional identities." *European Economic Review* 55(8):1046–1071.

Bisin, Alberto and Thierry Verdier. 2000. "A model of cultural transmission, voting and political ideology." *European Journal of Political Economy* 16(1):5–29.

Bisin, Alberto and Thierry Verdier. 2001. "The economics of cultural transmission and the dynamics of preferences." *Journal of Economic Theory* 97(2):298–319.

Bisin, Alberto and Thierry Verdier. 2010. The economics of cultural transmission and socialization. Technical report National Bureau of Economic Research.

Bowles, Samuel and Herbert Gintis. 2011. *A Cooperative Species: Human Reciprocity and its Evolution*. Princeton University Press.

Cantoni, Davide. 2015. "The economic effects of the Protestant Reformation: testing the Weber hypothesis in the German lands." *Journal of the European Economic Association* 13(4):561–598.

Carvalho, Jean-Paul. 2013. "Veiling." *The Quarterly Journal of Economics* 128(1):337–370.

Carvalho, Jean-Paul and Mark Koyama. 2013. Resisting Education. Technical report University Library of Munich, Germany.

Conquest, Robert. 1970. *The Nation Killers: The Soviet Deportation of Nationalities*. Macmillan London.

Froese, Paul. 2008. *The Plot to Kill God: Findings From the Soviet Experiment in Secularization*. University of California Press.

Greif, Avner. 2008. "The Normative Foundations of Institutions and Institutional Change." *Unpublished manuscript* .

Greif, Avner and Jared Rubin. 2014. "Endogenous Political Legitimacy: The English Reformation and the Institutional Foundations of Limited Government." *Working paper, Stanford University* .

Greif, Avner and Steven Tadelis. 2010. "A theory of moral persistence: Crypto-morality and political legitimacy." *Journal of Comparative Economics* 38(3):229–244.

Helfferich, Tryntje. 2009. *Thirty Years' War: An Anthology of Sources*. Hackett.

Huddy, Leonie, David O Sears and Jack S Levy. 2013. *The Oxford handbook of political psychology*. Oxford University Press.

Johnson, Noel D and Mark Koyama. 2013. "Legal centralization and the birth of the secular state." *Journal of Comparative Economics* 41(4):959–978.

Kula, Marcin. 2005. "Communism as Religion." *Totalitarian Movements and Political Religions* 6(3):371–381.

Laitin, David D. 1998. *Identity in formation: The Russian-speaking populations in the near abroad*. Vol. 22 Cambridge University Press.

Levi, Margaret. 1997. *Consent, dissent, and patriotism*. Cambridge University Press.

Levi, Margaret. 1999. "Death and taxes: extractive equality and the development of democratic institutions." *Democracy's Value* pp. 112–31.

Marshall, Richard H, Thomas E Bird and Andrew Blane. 1971. *Aspects of religion in the Soviet Union, 1917-1967*. University of Chicago Press.

Mendelson, Bert. 1975. *Introduction to topology*. Courier Corporation.

Nash, John. 1953. "Two-person cooperative games." *Econometrica* pp. 128–140.

Nexon, Daniel. 2009. "The struggle for power in early modern Europe." *Princeton University Press* .

North, Douglass C and Robert Paul Thomas. 1973. *The rise of the western world: A new economic history*. Cambridge University Press.

Northrop, Douglas. 2001. "Subaltern dialogues: Subversion and resistance in Soviet Uzbek family law." *Slavic review* pp. 115–139.

Nye, Joseph S. 1967. "Corruption and political development: A cost-benefit analysis." *The American Political Science Review* pp. 417–427.

Onorato, Massimiliano Gaetano, Kenneth Scheve and David Stasavage. 2014. "Technology and the Era of the Mass Army." *The Journal of Economic History* 74(02):449–481.

Petrosyan, Leon A. 2016. *Dynamic Games with Perfect Information*. Springer.

Philpott, Daniel. 2001. *Revolutions in sovereignty: how ideas shaped modern international relations*. Princeton University Press.

Rae, Heather. 2002. *State identities and the homogenisation of peoples*. Vol. 84 Cambridge University Press.

Ringmar, Erik. 2007. *Identity, interest and action: a cultural explanation of Sweden's intervention in the Thirty Years War*. Cambridge University Press.

Roberts, Michael. 1954. *The military revolution, 1560-1660*. Queen's University of Belfast.

Rokkan, Stein. 1999. *State formation, nation-building, and mass politics in Europe: the theory of Stein Rokkan: based on his collected works*. Clarendon Press.

Rubin, Jared. 2014. "Printing and Protestants: an empirical test of the role of printing in the Reformation." *Review of Economics and Statistics* 96(2):270–286.

Sambanis, Nicholas and Moses Shayo. 2013. "Social identification and ethnic conflict." *American Political Science Review* 107(02):294–325.

Schøyen, Øivind. 2011. *Legitimacy maximizing authorities*. Master thesis. The London School of Economics and Political Science.

Sutherland, Nicola M. 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, John. 1997. "The Demography of the Thirty Years War Re-revisited: Günther Franz and his Critics." *German History* 15(1):1–21.

Tilly, Charles. 1992. *Coercion, capital, and European states, AD 990-1992*. Blackwell Oxford.

Tilly, Charles and Gabriel Ardant. 1975. *The formation of national states in Western Europe*. Vol. 8 Princeton University Press.

Weingast, Barry R and Donald Wittman. 2008. *The Oxford handbook of political economy*. Oxford University Press.

Wilson, Peter Hamish. 2009. *The Thirty Years War: Europe's Tragedy*. Harvard University Press.

Chapter 2

Suspicious minds and views of fairness

Abstract

Do people with different views of what is fair attribute different intentions to actions? As one typically cannot observe the strength or type of people's fairness view or their material incentives, inferring intentions from actions is a multidimensional inference problem. I study this problem by investigating intention attribution in a setting where the monetary incentives are easily identifiable: a redistribution choice of voting for either full redistribution or no redistribution. Individuals with above median incomes will have monetary incentives for no redistribution, and individuals with below median incomes will have monetary incentives for full redistribution. In a novel experimental design, participants predict how likely other participants' redistribution vote is motivated by selfishness. I find that participants are significantly more likely to attribute a no redistribution vote to selfishness if they view redistribution as fair. I define this effect, attributing actions not adhering to one's own fairness view to selfishness, as suspicious attribution. I develop a theory of intention attribution to show how suspicious attribution can be explained by two other findings from the experiment: participants underestimate how many have a fairness view that differ from their own, projection bias, and overestimate the selfishness of participants with other fairness views, out-group stereotypes. My results and the idea of suspicious attribution can contribute to explaining polarization of attitudes and how prosocial individuals legitimize engaging in group conflicts.

1. Introduction

Behaviour that does not adhere to one's morality can either be attributed to another morality, or to self-interest.[1] Moralities other than one's own can often be hard to understand or empathize with (Haidt, 2007). Furthermore, empathizing with other moralities implies acknowledging objections to one's convictions, something people often have little incentive or interest in doing (Piketty, 1995). Thus, self-interest is often attributed to behaviour not adhering to one's own morality. I define this type of intention attribution, attributing behaviour not adhering to your own morality to selfishness, as suspicious attribution.

Understanding how differences in moralities affect our interpretation of intentions is central to understanding how polarization of attitudes occur (Haidt, 2012). Suspicious attribution will distort perceptions of how many people have selfish intentions in heterogeneous societies and increase the legitimacy of prosocial types to choose conflict over dialogue with out-groups. A recent example of such a process is the polarization of American politics; since the 1990s supporters of the Republican and Democratic parties have found less common ground. Both sides interpret the intentions of the other party with increasing suspicion, causing them to support increasingly off-centre candidates. This has led to a vicious cycle of polarization (Haidt, 2012) and political deadlock (Binder, 2015). Understanding the role of suspicious attribution is central to addressing the micro foundation of these processes.

I study suspicious attribution of voting in a simplified redistributive game: players vote for either full or no redistribution after observing their own income. In this simplified game of redistribution, the half of the players with above median incomes

[1] A morality can be understood as a vector of beliefs and values that are internalized and embedded in the person; examples are political ideologies, and religious or ethnic identities.

will have incentives for no redistribution, while the other half have incentives for full redistribution. Views of what constitutes fair redistribution also vary, e.g., Republicans and Democrats. Furthermore, it has been shown in experimental settings that where participants have no incentive in the outcomes, people hold mutually excluding fairness views about redistribution (Cappelen et al., 2007; Roemer, 2009). Some find it fair to redistribute earnings to compensate income differences arising from luck, effort or performance, hereby referred to as egalitarians, others prefer differences in earnings to be reflected in income differences, hereby referred to as libertarians.[2] Hence, simplified redistribution offers an ideal setting to study suspicious attribution; material incentives and fairness views prescribe predictable opposite alternatives, creating ambiguity of intentions that can easily be disentangled.

By definition, suspicious egalitarians will attribute no redistribution votes to selfishness, e.g., monetary incentives. In my redistributive game, this implies egalitarians will infer that no-redistribution votes are cast by selfish people with above median incomes. Similarly, a suspicious libertarian will attribute a vote for redistribution to be motivated by selfishness of people with below median income. Suspicious attribution of redistributive votes can arise from two types of beliefs about the type of voters who casts votes not adhering to one's fairness view. An egalitarian can infer that a no-redistribution voter is a selfish type of egalitarian with above median income; that is, a person with the same fairness view as his own, but who choose to vote according to monetary incentives rather than according to fairness view. Alternatively, an egalitarian can infer that the voter is indeed a libertarian; however, he might also believe that most libertarians are selfish types, i.e., mainly motivated

[2] My experimental design is not dependent on whether people's view conform to these labels; i.e., all that is assumed is that either participants find it fair to redistribute, or not redistribute.

by material incentives. I show that the first type of inference arises from projection bias in fairness views –inflated beliefs about how many people share your fairness view –while the latter arises from out-group stereotypes –biased beliefs about how many people with opposing fairness views are selfish types. These mechanisms are formally outlined in my theoretical framework of suspicious attribution.

In a novel laboratory experiment design I test for suspicious attribution of redistributive choices. Participants are randomly assigned either the role of predictor or the role of worker. Workers complete tasks, and earn money according to whether their output is above or below the median worker. Workers then vote for full or no redistribution after observing how large their earnings are compared to the median worker. Predictors are paid according to the accuracy of their estimate of how workers casting different redistribution votes have scored on a selfishness measure. The selfishness measure is the worker's average number of self-reported favourable coin tosses, in a setting where the worker is paid for every favourable coin toss and is not monitored. Implicitly, my design rests on the following assumption: if the predictor has a suspicious attribution of a worker's redistributive choice, the predictor will find it more likely that the worker will lie about the number of favourable coin tosses to increase his payment. Predictors also estimate how many workers hold each fairness view, how many workers are selfish types and how many workers are selfish types contingent on the workers' fairness views. To show suspicious attribution, a redistributive vote must be taken as a sign of being a selfish type, and not a sign of membership of another social group. Thus, the experiment was conducted on a sample with few or no salient social group cleavages.[3]

[3]Minimal group distinctions have been shown to change social inference from projection bias to stereotyping (Alicke, Dunning and Krueger, 2005).

The paper offers three main findings. *First*, I find suspicious attribution bias by egalitarians of no-redistribution votes; egalitarians predict a significantly higher probability of workers voting for no redistribution as being selfish types than libertarians. *Second*, I find that both libertarians and egalitarians display projection bias in fairness views; they overestimate the prevalence of their own fairness view. *Third*, I find that egalitarians have out-group stereotype bias against libertarians; their estimates of the prevalence of selfish types among libertarians is upward biased, and are significantly higher for egalitarian predictors compared with libertarian predictors. I find supportive evidence linking the three main findings for no-redistribution votes; the predictors' suspicious attribution bias is significantly correlated with their out-group stereotypes and prevalence estimates of fairness view.

This paper relates to several strands of literature. I link the literature on moral reasoning (Haidt, 2012), social cognition (Alicke, Dunning and Krueger, 2005) and fairness views (Cappelen et al., 2007; Roemer, 2009) by developing a specific model of inference of intentions behind distribution choices under projection bias and out-group stereotypes. Furthermore, the paper is novel in testing intention attribution behind redistributive choices in an experiment and finding supportive evidence of suspicious attribution. Projection bias was first demonstrated by Ross, Greene and House (1977), and is subsequently shown to hold across a large number of fields and situations (Blanco et al., 2014; Rubinstein and Salant, 2016; Bushong and Gagnon-Bartsch, 2016); I show projection bias also holds for fairness views. Out-group stereotypes have been previously found; people attribute negative intentions to people with differing opinions in general (Reeder et al., 2005), and exaggerate differences connected to political opinions in particular (Graham, Nosek and Haidt, 2012). These theories focus on perceived stereotypes of political attitudes, while I study predictions of intentions behind actions and the fairness view of the predictor. The experiment

applies the approach of using choices made by a party with no incentives in the outcomes, to signify fairness views first used by Harsanyi (1962). My experiment uses a coin flipping task as a measure of selfishness, previously applied as an unobtrusive measure of dishonesty (Cohn, Fehr and Maréchal, 2014).[4] My experiment uses multiple predictions of behaviour, selfishness and fairness views to disentangle participants' beliefs about intentions. This approach is, to the author's knowledge, novel. The theoretical model relates to recent work in social learning which investigated the complexities of inference under projection bias (Gagnon-Bartsch, 2017; Madarász, 2015). I develop a simple novel theory of intention attribution under three dimensions of uncertainty: monetary incentives, fairness views and selfishness.

This paper proceeds as follows. In part two, a game of attributing intentions to redistribution votes is presented. I use this game to demonstrate the proposed model of intention attribution. In part three, the experimental design is presented, then the results from the main experiment are presented in part four, before the paper concludes in part five. Appendixes contain proofs, experimental instructions and some further analysis.

2. A theory of intention attribution for a redistributive game

I analyze intention attribution of redistributive votes under projection bias and out-group stereotypes. I first present a simple redistributive game, before analyzing the choices of the workers and the beliefs of the predictors. I then show how suspicious attribution arises from projection bias and out-group stereotypes. The game corresponds to my experiment where I elicit all relevant parameters from the theory.

[4] In Appendix 4, I report results from an study that show egalitarians and libertarians have the same view of lying as a selfish action.

Proof of the propositions and the theorems are in Appendix 1.

2.1. A redistributive game

The game has two types of players: workers and predictors. The agents play the following game.

Stage 0. Nature randomly draws an odd number of workers and predictors. Players' social preferences differ along two dimensions: type, which can either be egalitarian or libertarian, $m \in \{\xi, \lambda\}$, and strength, which can either be weak or strong. A portion $\alpha \in (0,1)$ of the population is libertarian and $1-\alpha$ is egalitarian. A portion $\mu_m \in (0,1)$ of players with fairness views m have weak social preferences, and are referred to as selfish types, S, while $1-\mu_m$ have strong social preferences and are referred to as non-selfish types, NS.

Stage 1. Worker i undertakes a hidden action $a_i \in \{a_{ns}, a_s\}$ which is either selfish or unselfish. The worker receives a strictly positive amount of money if he chooses the selfish action, and nothing otherwise.

Stage 2. Worker i draws an output $o_i \in \{o_l, o_h\}$ which is either high, o_h, or low, o_l, with equal probability $\frac{1}{2}$, where $o_l < o_h$.

Stage 3. The worker casts a vote v_i for either full redistribution, v_r, or no redistribution, v_{nr}; $v_i \in \{v_r, v_{nr}\}$. The outcome is decided by a simple majority vote. If the full redistribution option receives a majority, everyone is paid an amount $\overline{y}$. If the no-redistribution option receives a majority, high-output workers receive an amount y_h and low-output workers receive an amount y_l, where $y_h > \overline{y} \equiv \frac{y_h+y_l}{2} > y_l \geq 0$.

Stage 4. Predictors observe the structure of the game, the support for distributions of types and one worker's vote. He does not observe the worker's output, the worker's fairness view or whether the worker performed the selfish action. A predictor receives a strictly positive amount of money if the predictor correctly predicts the exact probability that the worker has chosen a selfish action conditional on his vote.

As predictors only get paid for exact reporting, and I assume all players believe that the probability distributions for types are single-peaked, the game equilibrium is given by workers voting according to their type and predictors honestly reporting their beliefs.

2.2. Workers' choice

I now analyze the workers' choice of vote, and selfish or non-selfish actions. I first define what actions adhere to which fairness views, and make assumptions about the choices of selfish and non-selfish types. By definition, egalitarians find the outcome that all workers should receive $\overline{y}$ fair, while libertarians find the outcome that workers should receive y_h if they draw high output o_h, and y_l if they draw low output o_l,

fair.[5] As voting for redistribution, v_r, always makes the outcome that all workers are paid $\overline{y}$ more likely to occur, I define this vote, v_r, as adhering to the egalitarian fairness ideal. Likewise, casting a vote for no redistribution, v_{nr}, is defined as an action adhering to the libertarian fairness ideal. Both egalitarian and libertarian workers find it fair to perform the non-selfish action and unfair to perform the selfish action; i.e., the selfish action adheres to none of the fairness views and the non-selfish action adheres to both. Furthermore, I assume the following about selfish and non-selfish types.

Assumption 1. Choices of selfish types and non-selfish types

A non-selfish type i casts the vote v_i and chooses the action a_i that adheres to his fairness view. A selfish type j casts a vote v_j and chooses the action a_j that maximizes his material incentives.

[5] The defined fairness views can grow out of different beliefs about whether output at stage 2 in the game reflects a choice of effort or luck. It can also arise from differences in attitudes about what constitutes legitimate sources of equality. Roemer (2009) defines a *strict egalitarian* as a person who believes an equal distribution is a fair outcome regardless of the source of inequality, a *choice egalitarian* as a person who finds it fair to let income differences reflect effort differences but seek to redistribute inequality arising from luck, while a *libertarian* will never redistribute income differences arising from luck or effort. Assuming beliefs in our game do not vary, and both types believe the output draw is random, the egalitarian type can be a choice or strict egalitarian believing that workers should be held accountable for choices of effort. Assuming both libertarians and egalitarians believe the draw of output is not random, but a reflection of a choice of effort, the egalitarian must be a strict egalitarian and the libertarian can be either a choice egalitarian or a strict libertarian. Finally, allowing the beliefs about the draw of output to vary, the egalitarian type can be either a strict or a choice egalitarian believing that the draw is not random; the libertarian type can be a choice egalitarian believing the draw is not random, or a strict libertarian.

Selfish types, independent of fairness views, choose the selfish action a_s; non-selfish types choose the non-selfish action a_{ns}. From $y_h > \overline{y} > y_l$, it follows that o_l workers have material incentives to vote v_r, and o_h workers have material incentives to vote v_{nr}. Assumption 1, the redistribution mechanism and the definition of the fairness views, imply a worker votes according to the drawn output, strength and type of fairness view, as summarized in Table 1.

The outcome of the vote is dependent on the distribution of types. The distribution of votes for redistribution is given by: $P(v_r) = [\frac{\#v_r}{\#v_r + \#v_{nr}}] = \frac{1}{2}[\mu_\lambda \alpha + \mu_\xi(1-\alpha)] + (1-\mu_\xi)(1-\alpha)$. Note that, if everyone is selfish, $\mu = 1$, the prevalence of fairness views, α, does not matter for the vote outcome. If all workers are selfish types, $\mu_\xi = \mu_\lambda = 1$, then the distribution of votes only depends on the incentive structure for the actions, the redistribution mechanism and the draw of output, i.e., $P(o_l) = P(v_r) = \frac{1}{2}$. The prevalence of fairness views is increasingly important for the outcome of the vote if there is a high number of non-selfish workers; for $\mu_\xi = \mu_\lambda = 0$ the vote is proportional to the prevalence of fairness views, $P(v_r) = 1 - \alpha$. The aggregate vote share for different portions of selfish types and different fairness views is illustrated in Figure 1.

2.3. Predictors' beliefs and suspicious attribution

Suspicious attribution is defined as a belief that votes not adhering to your own fairness reflect a selfish motivation. I here analyze when a vote not adhering to the predictor's fairness view is taken as a signal that the worker is the selfish type. I first develop the Bayesian probability of a worker being a selfish type conditional on his vote, $P(S|v_i)$, describe when votes are uninformative of type, do comparative static

Choice Matrix Libertarian Workers

Portion α of the population

	High output(o_h)	Low output(o_l)
Selfish (μ_λ)	v_{nr}	v_r
Non-Selfish ($1-\mu_\lambda$)	v_{nr}	v_{nr}

Choice Matrix Egalitarian Workers

Portion $1-\alpha$ of the population

	High output(o_h)	Low output (o_l)
Selfish (μ_ξ)	v_{nr}	v_r
Non-Selfish ($1-\mu_\xi$)	v_r	v_r

Table 1: The choice of actions for the different types of persons. The prevalence of libertarians is α, while the prevalence of egalitarians is $1-\alpha$. The prevalence of selfish types among egalitarians is μ_ξ, and μ_λ among libertarians.

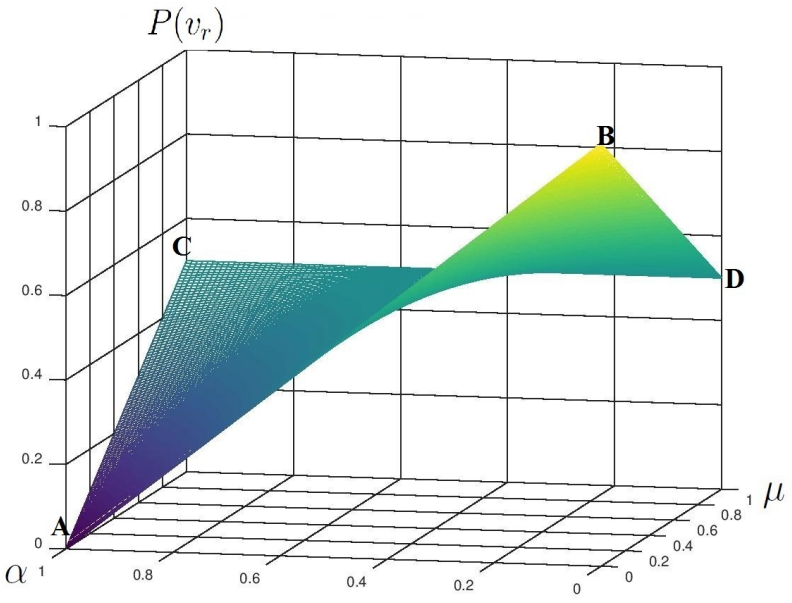


Figure 1: Theoretical distribution of the share of votes for redistribution, $P(v_r)$. The two horizontal axes give the share of people finding it fair not to redistribute, α, and the share of people prioritizing monetary incentives over fairness views, μ. The vertical axis gives the portion of votes for full redistribution, $P(v_r) = \frac{1}{2}[\mu_\lambda \alpha + \mu_\xi(1-\alpha)] + (1-\mu_\xi)(1-\alpha)$. In this graph, it is assumed that the portion of selfish types is equal for both groups of fairness views, i.e., the fairness view is uncorrelated with selfishness, $\mu_\lambda = \mu_\xi$. At point $A = \{\mu = 0, \alpha = 1, P(v_r) = 0\}$, all workers are non-selfish, hold a libertarian fairness view, and thus, all vote for no redistribution. At point $B = \{\mu = 0, \alpha = 0, P(v_r) = 1\}$, all workers hold an egalitarian fairness view and are non-selfish, thus, they all vote for redistribution. At the line from point $C = \{\mu = 1, \alpha = 1, P(v_r) = \frac{1}{2}\}$ to $D = \{\mu = 1, \alpha = 0, P(v_r) = \frac{1}{2}\}$, all workers are perfectly selfish, $\mu = 1$, and thus they all vote according to their output draw, which makes half of the votes for redistribution, $P(v_r) = \frac{1}{2}$. At the line between A and C everyone holds a libertarian fairness view, $\alpha = 1$, and as the portion of selfish types sinks from 1 to 0, $P(v_r)$ decreases from $\frac{1}{2}$ to 0. At the line between B and D everyone holds an egalitarian fairness view, $\alpha = 0$, and as the portion of selfish types sinks from 1 to 0, $P(v_r)$ increases from $\frac{1}{2}$ to 1.

on $P(S|v_i)$, and finally establish when suspicious attribution constitutes a bias.

The predictor observes what vote a worker has cast and reports the conditional probability that the worker is selfish. Votes for full redistribution are cast by selfish workers drawing low output, o_l, and non-selfish egalitarians. Votes for no redistribution are cast by selfish workers drawing high output, o_h, and non-selfish libertarians. The predictor's perceived probability that a worker is a selfish type after observing a vote $v_i \in \{v_r, v_{nr}\}$, denoted as $P(S|v_i)$, is as follows.

$$P(S|v_i) = \frac{\text{Number of selfish types casting } v_i}{\text{Number of selfish types casting } v_i + \text{ Number of non-selfish types casting } v_i} \tag{2.1}$$

Applying Bayes rule, I obtain the following expression for (2.1).

$$P(S|v_i) = \frac{P(v_i|S)P(S)}{P(v_i)} \equiv E[\mu_{v_i}] \tag{2.2}$$

I assume that predictors believe Assumption 1 is true: the incentive structure and preferences are assumed common knowledge; thus, predictors believe that workers vote according to Table 1. The predictors believe that a selfish worker votes according to his incentives and there is a $\frac{1}{2}$ probability that the worker draws o_l, and has incentives to vote v_{nr}. Thus, the probability of a selfish type voting for no redistribution is $P(v_{nr}|S) = \frac{1}{2}$. The probability of a worker being a selfish type is the sum of the probabilities of selfish types with both fairness views, $P(S)$, is $[\mu_\lambda\alpha + \mu_\xi(1-\alpha)]$. Non-selfish types vote for no redistribution if they are libertarian. Hence, the probability that a non-selfish worker votes for no redistribution is equal to the probability of a worker being a non-selfish libertarian, i.e., $P(v_{nr}|NS)P(NS) = 1(1-\mu_\lambda)\alpha$. By inserting this into (2.2), I attain the following.

$$E[\mu_{nr}] = \frac{\frac{1}{2}[\mu_\lambda\alpha + (1-\alpha)\mu_\xi]}{\frac{1}{2}[\mu_\lambda\alpha + \mu_\xi(1-\alpha)] + (1-\mu_\lambda)\alpha} \tag{2.3}$$

The term $[\mu_\lambda \alpha + (1-\alpha)\mu_\xi]$ is the weighted average rate of selfish types in the population, $\overline{\mu}$, which implies that (2.3) can be written as follows.

$$E[\mu_{nr}] = \frac{\overline{\mu}}{\overline{\mu} + 2(1-\mu_\lambda)\alpha} \tag{2.4}$$

Note from (2.4) that the expected probability of a worker being a selfish type contingent on him casting a non-redistribution vote increases as the number of selfish libertarians grows or the number of libertarians decreases. Replacing the portion of non-selfish libertarians with non-selfish egalitarians $2(1-\mu_\xi)(1-\alpha)$, in (2.4), I attain the probability of a selfish type conditional on observing a vote for redistribution.
First, I consider (2.3) under the special case where the predictor believes there is equal prevalence of each fairness view $\alpha = \frac{1}{2}$, and fairness views have equal portion of selfish types, $\mu_\xi = \mu_\lambda = \mu$. Note that, unless the effect of more selfish types among one of the fairness views exactly offsets the effect of prevalence, the special case of $\alpha = \frac{1}{2}$ and $\mu_\xi = \mu_\lambda$ is the only one where a vote is uninformative. This insight is presented as a theorem below.

Theorem of prior beliefs for uninformative votes

A predictor will not gain any information about whether a worker is selfish from a vote $v_i \in \{v_r, v_{nr}\}$ if he believes the following two conditions hold.
(i) There is an equal probability of a worker holding either fairness view, $\alpha = \frac{1}{2}$.
(ii) The portion of selfish types is equal for different fairness views, $\mu_\xi = \mu_\lambda$.

I now consider which prior beliefs will create which inferences when votes are informative of type. First, assume that predictors holds the belief that there is an equal share of selfish types among libertarians and egalitarians, $\mu_\xi = \mu_\lambda$. Then equation

(2.4) implies that a vote for no redistribution will be taken as a signal of being a selfish type, i.e., $E[\mu_{nr}] > E[\mu_r]$, if a majority of workers are egalitarians $\alpha < \frac{1}{2}$. A vote for redistribution will be taken as a signal of being a selfish type, $E[\mu_{nr}] < E[\mu_r]$, if a majority of workers are libertarians, $\alpha > \frac{1}{2}$. If I allow for predictors believing that there is correlation between being a selfish type and having a fairness view, i.e., $\mu_\xi \neq \mu_\lambda$, and assume predictors belive that there is equal prevalence of fairness views, $\alpha = \frac{1}{2}$, then whenever there are more selfish types among workers with fairness view m, $\mu_m > \mu_n$, the vote adhering to fairness view m will be taken as a signal of a selfish type. Second, I consider (2.3) under the special case where predictors believe that the portion of selfish types is equal for both fairness views, $\mu_\xi = \mu_\lambda = \mu$, while the expected α varies according to the fairness view of the predictor. Defining the predictors of fairness view m to hold the belief α^m, and denoting this expectation as E^m, I find the following inference for an egalitarian predictor observing a vote for no redistribution.

$$E^\xi[\mu_{nr}] = \frac{1}{1 + 2\alpha^\xi(\frac{1}{\mu} - 1)} \tag{2.5}$$

Note that the predicted probability of a worker considered to be selfish after casting a vote against redistribution decreases in the prevalence of libertarians, and increases in the prevalence of selfish types. The expression for a libertarian observing a redistribution vote is attained by replacing the term for non-selfish libertarians $\alpha^\xi(\frac{1}{\mu} - 1)$ with non-selfish egalitarians in (2.5). Furthermore, I assume that predictors overestimate the prevalence of their own fairness view, projection bias. Denoting the empirical portion of workers of libertarian fairness type as α^W, I define projection bias as follows.

Definition: Projection bias in fairness views implies $\alpha^{\xi} < \alpha^{W} < \alpha^{\lambda}$ (2.6)

It follows directly from the definition of projection bias that predictors will have a different conditional belief of selfishness for a vote not adhering to a predictor's view. Relaxing the assumption that there is an equal share of selfish types among holders of both fairness views, allowing for $\mu_{\xi} \neq \mu_{\lambda}$, and denoting a predictors belief of share μ_n of fairness view m, μ_n^m, I develop the following proposition.

Proposition 1. Intention attribution and projection bias in fairness views Take any two fairness views n and $m \in \{\xi, \lambda\}$, such that $m \neq n$. If holders of fairness views m and n have common beliefs about the portion of selfish types, $\mu_{\xi}^m = \mu_{\xi}^n$ and $\mu_{\lambda}^m = \mu_{\lambda}^n$, projection bias implies that a predictor with view m will cast a vote adhering to view n as a stronger signal of selfishness than a predictor with view n.

The intuition of the proof of the proposition is as follows. Considering (2.4), if there are no libertarians, $\alpha = 0$, then all workers voting for no redistribution are selfish types; a vote for no redistribution is a certain signal of being a selfish type. Increasing the number of libertarians increases the probability that a no-redistribution vote was cast by a non-selfish libertarian. A change in the prevalence of libertarian types α also changes the average selfishness $\overline{\mu}$ if there are different shares of selfish types amongst the different fairness views, $\mu_{\xi} \neq \mu_{\lambda}$. However, because average selfishness is a convex combination of selfishness for the two types of fairness views, this effect is always smaller than the change arising from changes in α. This implies that decreasing α always reduces the strength of the no-redistribution vote as a signal of

being a selfish type. Hence, under projection bias, egalitarians will always believe α to be lower than libertarians. Thus, for any common belief of the portion of selfish types amongst each group, $\mu_\lambda^m = \mu_\lambda^n, \mu_\xi^m = \mu_\xi^n$, egalitarians will take a vote for no redistribution as a stronger signal of selfishness than libertarians. The opposite holds for how libertarian predictors attribute intentions to a vote for redistribution.

I now analyze how a predictor's beliefs about the portion of selfish types change prediction about the probability of a worker being selfish a type contingent on observing the workers redistributive choice. Trivially, a predictor holding a belief that more workers, regardless of fairness view, are selfish types will give a higher probability that any vote v_i is committed by a selfish type; observe from (2.4) that for an egalitarian the portion of selfish types libertarians μ_λ will effect the probability of a selfish type casting a no redistribution vote both through the overall portion of selfish types and by the probability of the no redistribution vote being cast by a non-selfish type. Defining $\mu_{-j} - \mu_j$ as the out-group stereotype for a predictor of fairness view j, I can establish the following proposition on out-group stereotypes and non-adhering votes.

Proposition 2. Intention attribution and out-group stereotypes

Out-group stereotypes increase the difference between the predicted selfishness of workers casting votes not adhering and workers casting adhering to the fairness view of the predictor.

Following from Proposition 1 and 2 and defining a biased out-group stereotype as a deviation from the empirical portion of selfish types, I can establish the following theorem on suspicious attribution bias.

Theorem of suspicious attribution bias

Assume a predictor of fairness view m holds a rational or an upward biased belief about the portion of selfish types in his in-group, $\mu_m^m \geq \mu_m^W$. The predictor's belief that a vote not adhering to his fairness view is cast by a selfish type will be biased upward if either:

(i) the predictor has biased out-group stereotypes, $\mu_n^m > \mu_n^W$,

(ii) the predictor has projection bias in fairness views.

For a sufficiently downward biased view of in-group selfishness, i.e., $\mu_m^m < \mu_m^W$, the belief of a selfish type after observing a non-adhering vote might be unbiased even under projection bias and an upward bias in out-group selfishness. An example of how likely a vote for no redistribution is to be committed by a selfish type, as a function of μ_λ and α under $\mu_\xi = \frac{1}{2}$, is presented in Figure. 2.

3. Experimental Design

In this section, I present the experimental design and the participants' roles. Participants who signed up for the experiment were randomly allocated roles. Each participant had only one role, either as a *worker* or *predictor*. *Predictors* are paid according to the accuracy of their estimates of what votes and selfish choices workers made contingent on their vote and fairness view. The *worker* role is included in order to study predictions about concrete outcomes rather than hypothetical scenarios. By eliciting beliefs from predictors who have themselves not made the choices they are being questioned about, I reduce any risk of an upward bias in estimates because of predictors' norm seeking rationalization of choices. All participants are elicited for age, gender and whether they vote for political parties that actively pursue higher or lower levels of redistribution. A detailed overview of the experiment design and

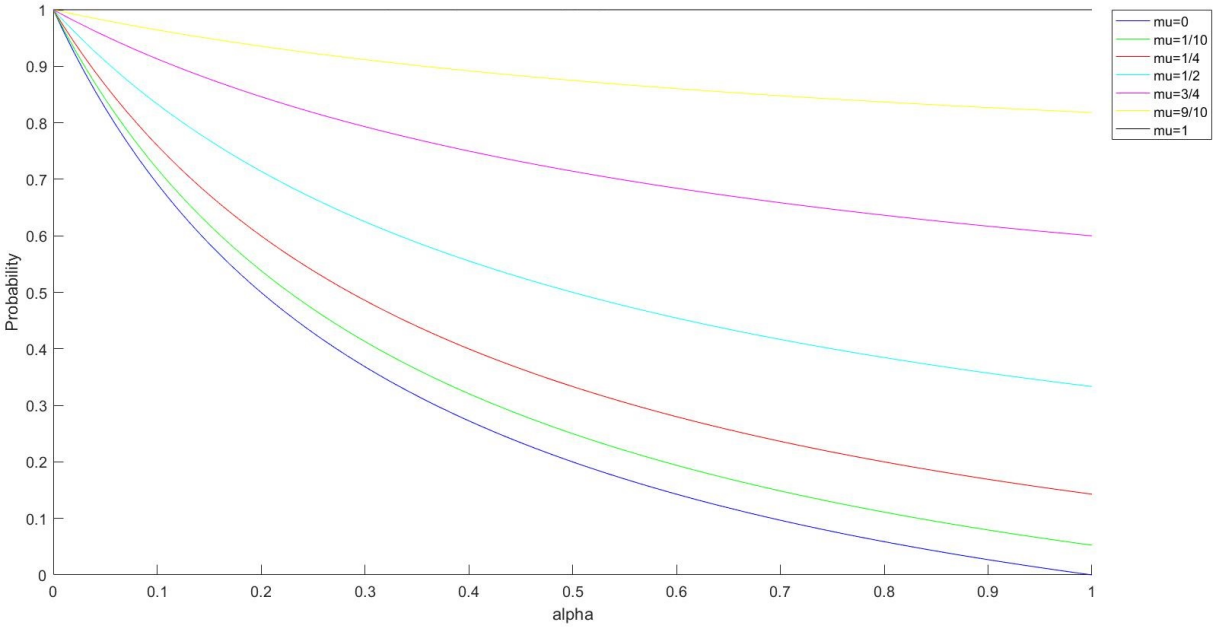


Figure 2: The vertical axis gives the conditional probability of a worker casting a no-redistribution vote being a selfish type, $E[\mu_{nr}] = \frac{\mu_\lambda \alpha + (1-\alpha)\mu_\xi}{[\mu_\lambda \alpha + \mu_\xi(1-\alpha)] + 2(1-\mu_\lambda)\alpha^\xi}$. The horizontal axis gives the perceived prevalence of libertarian types in the population α. Different lines reflect different perceptions of the portion of selfish types in the libertarian population, $mu \equiv \mu_\lambda$. The egalitarian selfishness prevalence, μ_ξ, is fixed at $\mu_\xi = \frac{1}{2}$. Note that the probability of a selfish type, given the observation of a vote for no redistribution, is decreasing in the prevalence of libertarians, α, and increasing in the prevalence of selfish types among libertarians, μ_ξ.

the roles in the experiment follow below.

Workers

The workers start by flipping a coin 10 times. Each worker is told to report the number of times the coin is flipped on the "tails" side, and that he will be paid four Norwegian kroner per tails outcome. There is no monitoring during the participants' coin flipping phase and the participants are aware of this. Consequently, the participants can misreport the number of times the coin landed on the "tails" side, to increase earnings. The coin flips by the workers are the empirical measure of selfishness.

Workers are then given 20 minutes to copy as many words as possible from a passage of text. After this, half of the workers with an above median number of correctly copied words among workers correctly identified earn a wage of 100 kroner, while the other half earn zero. Workers cast a vote for either full redistribution, so that workers are paid identically, or no redistribution, so that workers' payments equal their earnings. The option that receives a simple majority is implemented.

Predictors

First, predictors are presented with details of the worker's role and told that it was completed by participants who were randomly assigned to the role, and that workers were drawn from the same subject pool as themselves. Predictors are incentivized by the quadratic scoring rule to report their expected average number of reported coin flips by workers, contingent on their redistribution votes and fairness view.[6]

[6] The quadratic scoring rule is adopted following Blanco et al. (2014).

Predictors are asked about the group's average number of reported favourable coin flips for workers voting for redistribution $E^i[\mu_r]$, and workers voting against redistribution $E^i[\mu_{nr}]$. To elicit predictors' beliefs about any correlation between holding a particular fairness view and being a selfish type, I also elicit beliefs about the selfishness of worker participants according to their fairness view. To elicit predictors' beliefs about the correlation between fairness views and selfishness, I ask predictors about their beliefs about workers voting for or against redistribution which have no incentives over the outcome of the vote.[7] Predictors are incentivized to predict the selfishness of workers with no incentives over outcomes casting an un-incentivized vote for redistribution $E^i[\mu_\xi]$, against redistribution $E^i[\mu_\lambda]$ and how many workers with no incentives over the outcome of the vote find redistribution fair $E^i[1-\alpha]$.[8] Finally, predictors are asked about their own fairness view, $m_i \in \{\xi, \lambda\}$.

[7]To make these beliefs relate to concrete outcomes rather than hypothetical scenarios, I had a small group of six workers randomly selected from the same subject sample complete the worker role without actually working, i.e., these workers just voted for redistribution for other workers, flipped a coin 10 times and reported the number of favourable outcomes.

[8]An empirically observed predictor of fairness view m expectation of shares of selfish types μ_n is notated $E^m[\mu_n]$; while a theoretical predictor belief is notated μ_n^m. This is done to save notation in the theory section.

3.1. Overview of the experiment

Workers

Stage	1	2	3	4
	Workers flip a coin 10 times and report the number of favourable flips; μ_i.	Work Session 1: participants copy words from text.	All workers are informed whether his number of correctly copied words is above or below the median. (o_h/o_l)	Workers vote for either redistribution or no redistribution (r/nr) to determine first round payments; $(\overline{y})$ or $\{y_l, y_h\}$.

Predictors

Stage	1	2	3	4	5
	Predict parameters for workers: $E[\mu]$; $E[\mu_r], E[\mu_{nr}]$	Predict parameters for workers with no incentives in outcomes: $E[1-\alpha]$; $E[\mu_\xi], E[\mu_\lambda]$	Predictors flip a coin and report the number of favourable flips	Predictors report what option they find fair, which determines their type of fairness view $m \in \{\xi, \lambda\}$	Receive payment relative to prediction accuracy

Note: Overview of the different experiment stages for the two experimental roles, workers and predictors. The workers complete their stages before the predictors. Participants fill only one of two roles.

4. Descriptive statistics and results

A laboratory experiment with law students from the University of Bergen was conducted on September 3^{rd} 2015. The Faculty of Law at the University of Bergen has only one field of study, law. Consequently, sessions only involved students from the same field of study. If the data were drawn from, for example, the social science faculty where students have different specializations, i.e., sociology and economics, different redistributive votes would likely reflect stereotypes about different social groups. The sample size was 210 predictor participants and 18 worker participants. I conducted an online study on Amazon's Mechanical Turk platform (MTurk) of 99 American respondents to ensure that our selfishness measure is uncorrelated with ideological preferences, i.e., to confirm that the measure of selfishness is orthogonal to fairness views.[9] Details of all regression results in the main text are included in Appendix 2, Section 7.1. The results from the MTurk study are described in Appendix 4.

[9] I use which redistribution option predictors reported to find fair as a measure of their fairness view, rather than cast vote. About 15 per cent of the predictors voted for another redistribution alternative then what they reported to find fair. Except for projection bias in fairness views, all results were insignificant when applying predictors' behaviour, i.e., casting the third-party vote, as the measure of fairness view. This indicates that beliefs regarding other's intentions correlate with what predictors report as being fair, but has weaker correlation with what vote predictors believe is fair. The pre-experiment hypothesis and analysis plan are available in the online pre-analysis plan (Schøyen, 2017). This plan was posted online before the author had access to the data from the main experiment.

4.1. Descriptive statistics

The summary statistics of workers and predictors votes and how they scored on the measure of selfishness, reported the number of favourable coin tosses, is presented in Table 1. The majority of workers and predictors were egalitarians. The self-reported number of coin flips indicates that some participants exaggerated the number of favourable coin flips to earn more.

Table 1. Actual behaviour: group averages

	N	Egalitarians 1-α	μ_ξ	μ_λ
Predictors	210	56%	6.11	6.55
			(2.0)	(2.13)
	N	Votes for redistribution $P(v_r)$	μ_r	μ_{nr}
Workers	18	66%	5.5	6.33
			(1.62)	(1.63)

Note: Redistributive votes $P(v_r)$, fairness views $(1-\alpha)$, and reported favourable coin flips μ. The top line shows the average reported number of favourable coin flips out of 10 tosses, by predictors' fairness view and number of participants with an egalitarian fairness view. The bottom panel shows how many workers voted for redistribution and the average number of reported coin flips by their vote.

The summary statistics of predictors' expectations of worker behaviour are presented in Table 2. Predictors of both egalitarian and libertarian fairness views expected other participants to over-report the number of favourable coin flips. Workers voting for no redistribution were expected to over-report the number of coin flips to a

greater extent compared with workers voting for redistribution. As I show in the results section, the magnitudes and significances of the differences in expectations of selfishness between libertarian and egalitarian predictors increase when I compare differences within one predictor's estimates rather than with group averages.

Table 2. Predicted behaviour: group averages

	N	E[1-α]	E[μ_r]	E[μ_{nr}]	E[μ_ξ]	E[μ_λ]
Average predictions	210	60	5.42	6.83	5.75	6.58
		(24)	(1.29)	(1.70)	(1.63)	(1.29)
Average predictions by egalitarians (ξ)	126	67	5.33	6.94	5.72	6.72
		(21)	(1.25)	(1.69)	(1.26)	(1.63)
Average predictions by libertarians (λ)	84	50	5.56	6.65	5.79	6.37
		(25)	(1.33)	(1.70)	(1.38)	(1.61)

Note: The table shows the predicted prevalence of fairness views $E[\alpha]$ and predicted average reported favourable coin flips $E[\mu]$, among workers according to their votes $E[\mu_r]$, $E[\mu_{nr}]$ and according to their fairness view, $E[\mu_\xi]$, $E[\mu_\lambda]$.

4.2. Results

To study suspicious attribution, I focus on the within-predictor difference in estimated average difference in selfishness between workers voting for or against redistribution, $E^i[\mu_{nr} - \mu_r]$. While predictors of both fairness views predict that workers voting for no redistribution are more selfish, predictors with an egalitarian fairness view predict a larger difference between workers voting for no redistribution and for redistribution than libertarian predictors. The differences between egalitarian and

libertarian predictors' selfishness estimates according to workers' votes, $E^i[\mu_{nr} - \mu_r]$, and the actual difference, $\mu_{nr} - \mu_r$, is shown in Figure 3. The figure shows the average difference between predictors estimate of average selfishness for workers casting a vote for redistribution and against redistribution for egalitarian and libertarian predictors.

This finding can also be shown in a regression framework; I now turn to a regression analysis of the correlation between predictors' expected differences in workers' selfishness according to their votes and the predictors' fairness view. The binary variable D^{ξ} indicates a predictors fairness view; it is equal to 1 if a predictor finds the redistribution option the fair option and 0 else. I estimate the following.

$$E^i[\mu_{nr} - \mu_r] = \beta_0 + \beta_1 D^{\xi} + \epsilon_i \tag{4.1}$$

The difference in predictors' expectations of selfishness contingent on vote according to the predictors' fairness view β_1 in 4.1, is estimated to be 0.51 (0.063) while β_0 is estimated to be 1.093(0.000); egalitarians predict the difference in selfishness between workers voting for redistribution or no redistribution is about one-third larger than libertarians.[10] This is the main finding of the experiment: egalitarians consider a vote for no redistribution to be a significantly stronger signal of selfishness than libertarians, suggesting that the fairness view affects intention attribution. Based on this I establish the first result as follows.

Result 1: Participants' interpretation of intentions behind redistribution choices differ according to their fairness view.

[10]The probability of no effect given the observed estimate, p-value, are stated in parenthesis throughout the main text. The standard deviations of the estimates can be found in Section 7.1 in the appendix.

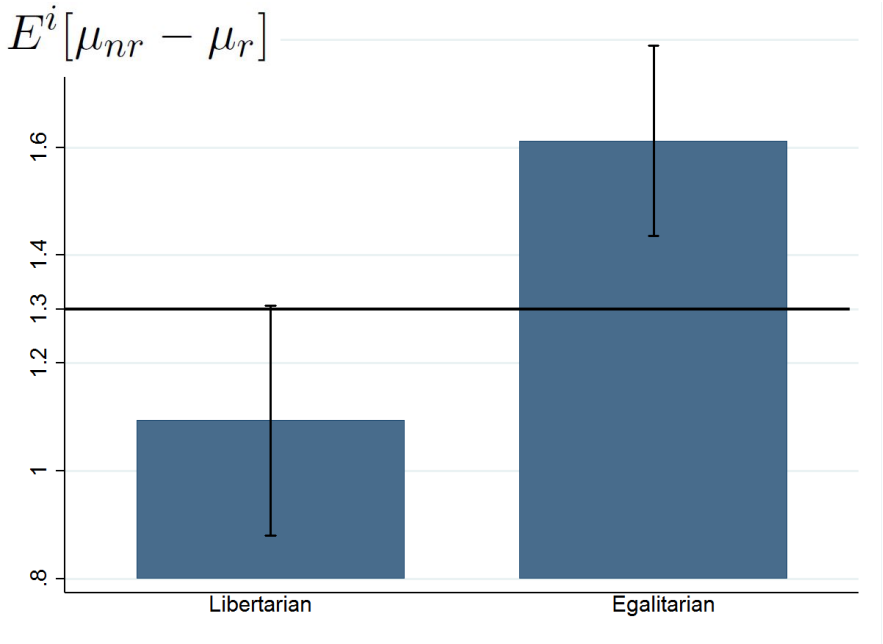


Figure 3: Differences in the predictors' reported expected average number of coin flips between workers voting for redistribution, full redistribution or no redistribution $E^m[\mu_{nr} - \mu_r]$, reported by predictors' fairness view $m \in \{\xi, \lambda\}$. The solid horizontal line the sample differences between workers voting for no redistribution or voting for redistribution $\mu_{nr} - \mu_r$, 1.3 coin flips. Egalitarians display significant suspicious attribution bias for no-redistribution votes. Libertarians have an insignificant suspicious attribution bias for distribution votes.

The differences in averages observed in Figure 3 imply that the suspicious attribution of no-redistribution votes among egalitarian predictors drives Result 1. In the theory section I have established that suspicious attribution can be driven by both projection bias in fairness views and out-group stereotypes. I now analyze whether predictors have projection bias and out-group stereotypes. I then analyze whether these beliefs drive Result 1.

As shown in Figure 4 the predictors displayed projection bias in fairness views; libertarian predictors reported they thought half the workers were libertarians, $E[1-\alpha^{\lambda}]=0.5$, while egalitarians thought about two thirds of workers were egalitarians, $E[1-\alpha^{\xi}]=0.67$. The actual prevalence among the 210 predictors was in between these estimates, $1-\alpha^{\xi}=0.56$; thus, both egalitarian and libertarian predictors overestimated the commonness of their fairness view and predictors of both fairness views displayed projection bias. This can be seen from Figure 4. Based on this, I establish my second result as follows.

Result 2: Projection bias in fairness views

I find out-group bias, measured by the difference between real and expected selfishness, for workers with an egalitarian or libertarian fairness view. This can be seen from Figure 5. Egalitarian predictors expected the difference in selfihness for workers casting un-incentivized votes against and for redistribution, $E^{i}[\mu_{\lambda}-\mu_{\xi}]$, to be equal to 1, while libertarian predictors expected this difference to be equal to 0.58. The difference for the sample $\mu_{\lambda}-\mu_{\xi}$ for the 210 predictors was 0.57. Hence, egalitarian predictors has an upwardly biased perception of the selfishness libertarian workers, while libertarian predictors has accurate beliefs. The difference between observed behavior and expected selfishness, according to the predictors' fairness view, is illustrated in Figure. 5.

I now turn to a regression analysis of the correlation between the differences in out-group stereotypes, measured by expected selfishness according to workers' fairness view, $E^{i}[\mu_{\lambda}-\mu_{\xi}]$, and predictors' fairness view, D^{ξ}.

$$E^{i}[\mu_{\lambda}-\mu_{\xi}]=\beta_{0}+\beta_{1}D^{\xi}+\epsilon_{i} \tag{4.2}$$

The effect of a predictor being egalitarian with regards to out-group stereotypes, β_1, is estimated at 0.41(0.11). Both egalitarians and libertarians expect libertarians to be more selfish, but egalitarians more so and more than the actual observed difference. Thus, I find that egalitarians have an out-group stereotype in their beliefs about the selfishhness of the libertarians, while I find no evidence of this for the libertarians. I establish the following as a result.

Result 3: Egalitarians have biased out-group stereotypes against libertarians.

I find that predictors' prevalence estimate of egalitarian fairness views among the workers, $E^i[1-\alpha]$, and out-group stereotypes against libertarians, $E^i[\mu^\xi - \mu^\lambda]$, are significantly positively correlated with attribution of selfishness to no-redistribution votes. I regress the difference in reported coin flips, $E^i[\mu_{nr} - \mu_r]$, on the prevalence estimates as follows.

$$E^i[\mu_{nr} - \mu_r] = \beta_0 + \beta_1 E^i[1-\alpha] + \epsilon_i \tag{4.3}$$

Expecting more participants to hold an egalitarian fairness view significantly correlates with predictions of the difference in selfishness between workers voting for no redistribution and redistribution; the effect on predictors' estimates of $E^i[\mu_{nr} - \mu_r]$ of one more participant holding an egalitarian fairness view out of 100 workers, $E^i[1-\alpha]$, is estimated at 0.95(0.093).

Out-group stereotypes against libertarians significantly correlated with suspicious attribution of no-redistribution votes, $E^i[\mu_{nr} - \mu_r]$. I measure out-group stereotypes as measured by differences in selfishness workers of different fairness views, $E^i[\mu_\lambda - \mu_\xi]$, as follows.

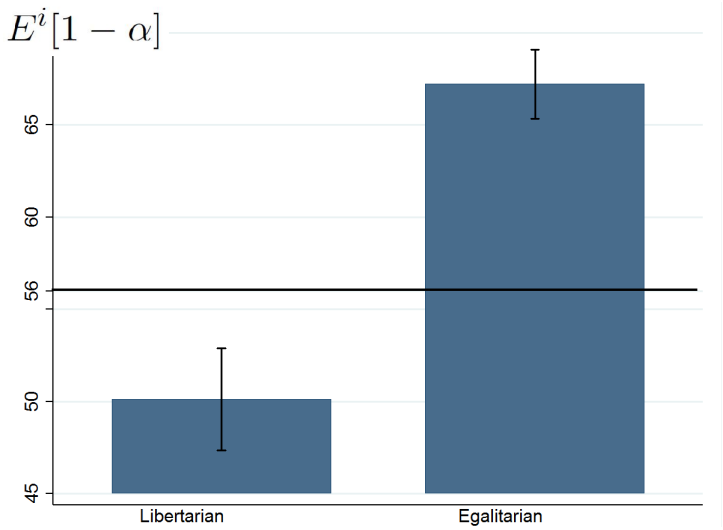


Figure 4: Distribution of predictors' estimates of how many workers find redistribution fair, $E^i[1-\alpha]$, reported by predictors' fairness view. The solid horizontal line indicates the sample value for the 210 predictors; $(1-\alpha) = 56$ per cent of predictors cast an un-incentivized vote that workers should redistribute.

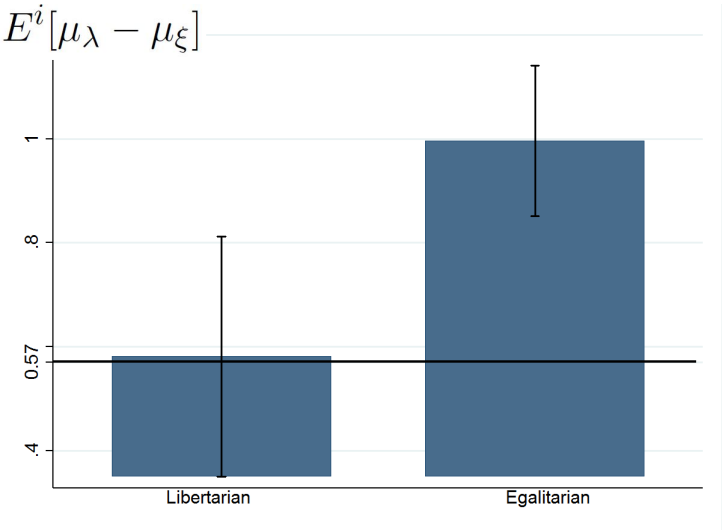


Figure 5: Difference in predictors' reported expected average number of coin flips between workers casting an un-incentivized vote for no redistribution or full redistribution, $E^i[\mu_\lambda - \mu_\xi]$, reported by predictors' fairness view. The solid horizontal line indicates true differences in reported coin flips between participants casting an un-incentivized vote for no redistribution or full redistribution in the sample of 210 predictors; $\mu_{nr} - \mu_r = 0.56$ coin flips.

$$E^i[\mu_{nr} - \mu_r] = \beta_0 + \beta_2 E^i[\mu_\lambda - \mu_\xi] + \epsilon_i \tag{4.4}$$

The effect of out-group stereotypes on suspicious attribution β_2 is estimated to be 0.40(0). Based on this, I establish my fourth and final result.

Result 4: Both projection bias and out-group stereotypes account for suspicious attribution of no redistribution votes.

Out-group stereotypes explain suspicious attribution to a larger degree than projection bias; regressing predictors expected differences in workers selfishness according to workers votes, $E^i[\mu_{nr} - \mu_r]$, on out-group stereotypes, $E^i[\mu_\lambda - \mu_\xi]$, (4.4), gives a higher explained variance R^2 than regressing on prevalence estimates,$E^i[1 - \alpha]$, (4.3). The correlation between differences in workers selfishness according to vote and prevalence estimates, becomes insignificant when including out-group stereotypes effects. The correlation between out-group stereotypes is robust to the inclusion of control variables; out-group stereotypes correlate with degree of suspicious attribution when the control variables are included. Details on this results along with additional regression results in shown in Section 7.2 of Appendix 2. In generall the results indicates that out-group stereotypes are the main driver of the suspicious attributions of no-redistribution votes.

Summary of main findings

Defining a bias as a deviation from the observed sample averages, my findings of fairness view are summarized in Table 4.

Table 4. Main findings: Differences in individual predictors.

	Egalitarians	Libertarians
Suspicious attribution bias	Significant bias	Insignificant Bias
Projection bias	Significant bias	Significant bias
Out-group stereotype bias	Significant bias	Accurate beliefs

Note: The table shows which biases are found in the beliefs of egalitarian and libertarian predictors.

Note that the theory section predicts that libertarians should have weaker suspicious attribution of no-redistribution votes if their beliefs about out-group stereotypes are accurate; my data are supportive of such a mechanism. Finally, note from Figure 3 and Figure 4 that egalitarian predictors believe that workers casting a vote for no redistribution are more selfish than workers with a libertarian fairness view. This difference in beliefs about workers casting a vote or holding a fairness view is also reflected in the level differences in Table 3. This finding is compatible with the theoretical framework of suspicious attribution; egalitarians predictors believe no-redistribution votes are cast by both selfish egalitarians and by libertarians, whom they believe there to be fewer of and to be more selfish than egalitarians.

5. Conclusion

I find evidence that people systematically interpret intentions behind redistribution choices differently according to their fairness views: egalitarian predictors have significant suspicious attribution bias of no-redistribution votes, whereas libertarian predictors have insignificant suspicious attribution bias of redistribution votes. The predictors were paid according to their accuracy and the setting encouraged them to

provide their best guess of the behaviour of fellow in-group members. The combination of material incentives and framing creates competition for accuracy and makes it reasonable to assume the measured effect does not reflect differences in participants' points of view; rather, the experiment shows supportive evidence of suspicious attribution among egalitarians. Furthermore, suspicious attribution in my experiment is correlated with projection bias and out-group stereotypes, which supports the premises of the proposed model of intention attribution.

Suspicious attribution is potentially crucial for understanding the maintenance of social group boundaries, especially why social groups do not erode in the face of heterogenous behaviour. Suspicious attribution can account for maintenance of social group boundaries as it enables people to sustain their belief in the superiority of their own fairness view. Let us assume that people did not have any suspicious attribution bias when considering the intentions of heterogenous behaviour of in-group members, where by in-group members are meant individuals considered to be relevant sources of social signals. In the absence of suspicious attribution bias, people's estimates of heterogeneity of fairness views would increase as they encounter heterogenous behaviour among in-group members. Heterogenous behaviour would not be attributed to incentive differences; the behavior would the be attributed to heterogenous social preferences. This would then lead people to question their own social preferences, or question the relevance of signals from in-group members as relevant social information. These two mechanisms would erode identification with social groups, but are avoided under suspicious attribution.

There are several theoretical and empirical questions arising from this work that can be addressed in future research. A promising theoretical extension is to analyze what actions can be equilibrium signals of low-selfishness, i.e., faithful attribution. This question can be analyzed by allowing workers to have preferences for identi-

fication as a non-selfish type, to investigate how material incentives, fairness views and prevalence of types, interact to allow different actions to signal non-selfishness in equilibrium. Such an extension of the theory can address questions of how sophistication of beliefs, i.e., second or third order beliefs of how others infer intentions, affect inference in real life. In particular, it seems promising to analyze the development of "cheap talk" political correctness or controversial stands as a costly signal of being a type concerned with outcomes. Furthermore, adding dynamic processes of updating can address how fast updating happens when incentives for actions change. In other words, when do actions that are antiquated in bringing about good outcomes linger as a signal of non-selfishness after incentives for actions have changed? This can help explain what determines whether a society is prone to conservative or progressive attitudes. Another possible extension is analyzing how much provision there will be of a particular type of action as a function of its ability to signal being a non-selfish type.

The experiment design can address a number of other empirical examples to investigate intention attribution in other settings. Finding suitable situations where ulterior motives or differences in internalized beliefs could explain differences in choice arising from salient cleavages other than political, such as ethnic, gender, national, religious or other social cleavages. One example could be religious practices. Assume that dedication to a faith and beliefs about the correct practice vary; do more orthodox believers suspiciously attribute the intentions of liberal statements? Another application of the model could be to understand how people attribute intentions to stated views on immigration policies. Assume people hold varying beliefs about the social desirability of open borders and varying degrees of willingness to contribute to social welfare; do people believing in the social desirability of an open border immigration policy have suspicious attribution of votes against open borders?

Suspicious attribution bias has troublesome implications for societies with large plurality of fairness views. The bias implies that larger heterogeneity of behaviour leads to more behaviour being attributed to selfishness. This will lead members of heterogenous societies to have a negative bias in their estimates of the number of non-selfish types. Moving beyond attribution of non-selfish intentions, to attribution of hostile intentions, the mechanism can contribute to cycles of distrust and eventual conflict, cycles that can arise even among rational actors (Acemoglu and Wolitzky, 2014). Hence, perhaps the most important line of future research is to investigate what can decrease our tendency towards suspicious attribution. This research can contribute to identifying interventions that can reduce prejudice and conflict (Paluck, 2012) and help to avoid traps of suspicious attribution.

Acknowledgments

I am grateful to FAIR Centre and the Department of Economics for financial and organizational support in conducting the experiment. I would like to thank Bertil Tungodden for excellent supervision and help. I would also like to thank Tristan Gangon-Bratch, Thor Øivind Jensen, Stefan Meissner, Tom Grimstvedt Meling, Jonas Tungodden, seminar participants at the Department of Strategy and Management at the Norwegian School of Economics, 2016 Meeting of the Norwegian Association of Economists at the Norwegian University of Science and Technology, and the OsloMet Economics Department for comments and encouragement. Finally, I would like to thank Xianwen Chen, Erik Eikeland, Vivienne Knowles, Karin Lillevold and Ida Kjørholt for research assistance. I gratefully acknowledge support from the Departments of Economics at the Norwegian School of Economics and Norwegian Research Council Grant Numbers 236995 and 262675.

6. Appendix 1: Proofs

Proof of theorem of no suspicious attribution

A predictor will not gain any information about whether a worker is selfish from any vote $v_i \in \{v_r, v_{nr}\}$ if he believes that the following two conditions hold.
(i) There is an equal probability of a worker holding either fairness view, $\alpha = \frac{1}{2}$.
(ii) The probability of a selfish type for all fairness views is equal to $\mu_\xi = \mu_\lambda$.

Proof: A non-informative vote implies that the probability of a selfish type is equal for both votes, i.e., $E[\mu_r] = E[\mu_{nr}]$. Thus, the theorem implies that if $\mu_\xi = \mu_\lambda$ and $\alpha = \frac{1}{2}$, then $E[\mu_r] = E[\mu_{nr}] = \mu'$. Substituting $\mu_\xi = \mu_\lambda = \mu'$ and $\alpha = \frac{1}{2}$ into (2.3) gives $E[\mu_{nr}] = \mu'$.
□

Proof of Proposition 1. Intention attribution and projection bias in fairness views

Take any two fairness views n and $m \in \{\xi, \lambda\}$ such that $m \neq n$. If holders of fairness views m and n have common beliefs about the portion of selfish types, $\mu_\xi^m = \mu_\xi^n$ and $\mu_\lambda^m = \mu_\lambda^n$, projection bias implies a predictor with view m taking a vote adhering to view n as a stronger signal of selfishness than a predictor with view n.

Proof: The definition of projection bias, (2.6), implies that the proposition holds if a vote not adhering to the predictor's fairness view is always taken as a stronger signal of selfishness for lower prevalence of opposing fairness view, i.e., $\frac{\partial E^\xi[\mu_{nr}]}{\partial \alpha} < 0$ and $\frac{\partial E^\lambda[\mu_r]}{\partial (1-\alpha)} < 0$ must hold for any μ_ξ, μ_λ.
I now show that $\frac{\partial E^\xi[\mu_{nr}]}{\partial \alpha} < 0$ holds for any $\mu_\xi^\xi, \mu_\lambda^\xi$. From (2.3) I attain that requiring

$\frac{\partial E^{\xi}[\mu_{nr}]}{\partial \alpha} < 0$ implies $[\mu_{\lambda}^{\xi}\alpha + \mu_{\xi}^{\xi}(1-\alpha^{\xi})] > [\mu_{\xi}^{\xi} - \mu_{\lambda}^{\xi}]$. This implies that the weighted average of beliefs of the prevalence of selfish types for both fairness views must be larger than the difference in the prevalence of selfish types between the two groups. As average selfishness is a convex combination of selfishness of the two groups, and I have assumed that the support of the prevalence of selfish types is $\mu_i \in (0,1)$, this strict inequality always holds. Hence, $\frac{\partial E^{\xi}[\mu_{nr}]}{\partial \alpha} < 0$ holds for any μ_{ξ}, μ_{λ}. $\frac{\partial E[\mu_r]}{\partial (1-\alpha)} < 0$ follows mutatis mutandis.

□

Proof of Proposition 2. Intention attribution and out-group stereotypes Out-group stereotypes increase the difference between the predicted selfishness of workers casting votes not adhering and workers casting adhering to the fairness view of the predictor.

Proof: Out-group stereotypes are defined as $\mu_{-j} - \mu_j$. Thus the proposition implies that the difference in perceived selfishness between votes for no redistribution and redistribution should increase in the difference in the portion of selfish libertarians and egalitarians; $\frac{\partial E[\mu_{nr} - \mu_r]}{\partial [\mu_{\lambda} - \mu_{\xi}]} > 0$. It follows from the probability of a no redistribution vote being cast by a selfish type, (2.4), and the equivalent expression for redistribution votes that

$$E[\mu_{nr} - \mu_r] = \overline{\mu}[\frac{1}{\overline{\mu} + 2(1-\mu_{\lambda})\alpha} - \frac{1}{\overline{\mu} + 2(1-\alpha)(1-\mu_{\xi})}]$$

, which always increases in μ_{λ}, decraces in μ_{ξ} and thus grows in $[\mu_{\lambda} - \mu_{\xi}]$. The proof for the difference between a redistribution or no-redistribution vote and $[\mu_{\xi} - \mu_{\lambda}]$ follows mutatis mutandis.

□

Proof of Theorem of Suspicious Attribution Bias

Assume a predictor of fairness view m holds a rational or an upward biased belief about the portion of selfish types in his in-group, $\mu_m^m \geq \mu_m^W$. The predictor's belief that a vote not adhering to his fairness view is cast by a selfish type will be biased upward if either:

(i) the predictor has biased out-group stereotypes, $\mu_n^m > \mu_n^W$,

(ii) the predictor has projection bias in fairness views.

Proof: For an egalitarian predictor the belief that a worker is a selfish type after observing a vote not adhering to his fairness view is $E^{\xi}[\mu_{nr}]$. Assume the empirical portions are $\{\mu_\lambda^W, \mu_\xi^W, \alpha^W\}$ and an egalitarian predictor holds some beliefs $\{\mu_\lambda^\xi, \mu_\xi^\xi, \alpha^\xi\}$ such that $\mu_\lambda^\xi > \mu_\lambda^W, \alpha^\xi < \alpha^W, \mu_\xi^\xi \geq \mu_\xi^W$. Part (i) of the theorem follows from that it is assumed that $\mu_m^m \geq \mu_m^W$, out-group stereotypes are defined as $\mu_\lambda - \mu_\xi$ and $\frac{\partial E^{\xi}[\mu_r]}{\partial \mu_\xi^\xi} > 0$ which follows from (2.3). Thus $E^{\xi}[\mu_{nr}]$ will be upward biased if $\mu_\lambda^\xi > \mu_\lambda^W$. Claim (ii) follows from $\frac{\partial E^{\xi}[\mu_{nr}]}{\partial \alpha^\xi} < 0$, as follows from the proof of Proposition 1. That a libertarian will have biased beliefs of a selfish type, $E^{\lambda}[\mu_r]$, according to (i) and (ii) follows mutatis mutandis.

□

7. Appendix 2: Further data

7.1. Main regression results

Result 1: Participants' interpretation of intentions behind redistribution choices differ by fairness view. Predictors' estimated difference in selfishness between workers voting for no redistribution or full redistribution is higher for egalitarian predictors (4.1) :$E^i[\mu_{nr} - \mu_r] = \beta_0 + \beta_1 D^{\xi} + \epsilon_i$.

Table A 1. Suspicious attribution and fairness view

	Dependent variable:
	$\mathrm{E}^i[\mu_{nr} - \mu_r]$
D^{ξ}	0.518*
	(0.278)
Constant	1.093 ***
	(0.216)
Observations	210
R^2	0.0165
Adjusted R^2	0.0117
Residual std. error	807.427 (df = 208)
F statistic	3.48 (df = 1; 208)
Note:	*p<0.1; **p<0.05; ***p<0.01

Result 3: Biased out-group stereotypes against libertarians among egalitarians.

Egalitarian predictors estimate higher differences between egalitarian and libertarian selfishness: (4.2) :$E^i[\mu_\xi - \mu_\lambda] = \beta_0 + \beta_1 D^\xi + \epsilon_i$.

Table A 2. Out-group stereotypes and fairness view

	Dependent variable:
	$E^i[\mu_\xi - \mu_\lambda]$
D^ξ	0.414
	(0.258)
Constant	0.581***
	(0.201)
Observations	210
R^2	0.012
Adjusted R^2	0.007
Residual std. error	698.886 (df = 208)
F statistic	2.57 (df = 1; 208)
Note:	*p<0.1; **p<0.05; ***p<0.01

Result 4: Both projection bias and out-group stereotypes account for suspicious attribution of no-redistribution votes.

Projection bias predicts differences in predictors' estimates of selfishness between workers voting for high and no redistribution: (4.3): $E^i[\mu_{nr} - \mu_r] = \beta_0 + \beta_1 E^i[1 - \alpha] + \epsilon_i$.

Table A 3. Correlation between suspicious attribution and projection bias.

	Dependent variable:
	$\mathrm{E}^i[\mu_{nr} - \mu_r]$
$\mathrm{E}^i[1 - \alpha]$	0.948^{**}
	(0.093)
Constant	0.833^{***}
	(0.024)
Observations	210
R^2	0.013
Adjusted R^2	0.008
Residual std. error	809.825 (df = 208)
F statistic	2.85 (df = 1; 208)
Note:	*p<0.1; **p<0.05; ***p<0.01

The out-group stereotype predicts differences in predictors' estimates of selfishness between workers voting for high and no redistribution: (4.4) $E^i[\mu_{nr} - \mu_r] = \beta_0 + \beta_2 E^i[\mu_\lambda - \mu_\xi] + \epsilon_i$.

Table A 4. Correlation between suspicious attribution and out-group stereotypes.

	Dependent variable:
	$\mathrm{E}^i[\mu_{nr} - \mu_r]$
$\mathrm{E}^i[\mu_\lambda - \mu_\xi]$	0.396^{***}
	(0.069)
Constant	1.077^{***}
	(0.139)
Observations	210
R^2	0.131
Adjusted R^2	0.131
Residual std. error	709.900 (df = 208)
F statistic	32.53 (df = 1; 208)
Note:	*p<0.1; **p<0.05; ***p<0.01

7.2. Details of Result 4: Both projection bias and out-group stereotypes account for suspicious attribution of no-redistribution votes.

I show here that Result 4 in the results section is robust. I show how explained variance R^2 and significance change with the inclusion of different variables that explain predictors' expected differences between the selfishness of workers voting for redistribution or no redistribution, $E^i[\mu_{nr} - \mu_r]$. I define $E^i[\mu]$ as the expectation of a general participant's reported number of coin flips, and D_{female} as a binary variable which is one if the predictor is female. Finally, I define the variable D_{RParty} and $D_{NRParty}$ as a binary variable indicating whether the participant reported voting for a political party actively pursuing redistribution at the last national election. I regress the difference in predictors' expected differences in average coin flips for workers voting for and against redistribution, on both their prevalence estimate of egalitarians, their out-group stereotypes and with and without control variables.

$$E^i[\mu_{nr} - \mu_r] = \beta_0 + \beta_1 E^i[1 - \alpha] + \beta_2 E^i[\mu_\lambda - \mu_\xi] + \epsilon_i \tag{7.1}$$

$$E^i[\mu_{nr} - \mu_r] = \beta_0 + \beta_1 E^i[1 - \alpha] + \beta_2 E^i[\mu_\lambda - \mu_\xi] + \beta_3 E^i[\mu] + \beta_4 D_{female} + \beta_5$$

$$\mathrm{D}_{RParty} + \beta_6 \mathrm{D}_{NRParty} + \epsilon_i \tag{7.2}$$

Table A 1 compares the explained variance, R^2, from the regressions and the coefficients with the from regressing $E^i[\mu_{nr} - \mu_r]$ on only the predictors prevalence estimate of egalitarian predictors $E^i[1 - \alpha]$ (4.3), and out-group stereotypes $E^i[\mu_\lambda - \mu_\xi]$ (4.4):

Table A 5. Regression results from estimating (4.3)**,** (4.4)**,**(7.1) **and** (7.2)

Table 2: Correlation of suspicious attribution of no-redistribution votes $E^i[\mu_{nr} - \mu_r]$.

Regression	$E^i[\alpha]$	$\mathrm{E}^i[\mu_\lambda - \mu_\xi]$	R^2
(4.3)	0.95*	·	0.013
	(0.56)		
(4.4)	·	2.68*	0.013
		(1.39)	
(7.1)	0.67	0.39***	0.142
	(0.52)	(0.07)	
(7.2)	0.70	0.38***	0.142
	(0.52)	(0.07)	

Table 3: *, P-value < 0.10, ***, P-value < 0.01.

Note: The table shows how predictors' expected prevalence of egalitarians, $E^i[\alpha]$, and out-group stereotypes correlate with their suspicious attribution of votes against redistribution, $E^i[\mu_\lambda - \mu_\xi]$.

This evidence suggests that both projection bias and out-group stereotypes contribute to the interpretation of intentions, but that out-group stereotypes are the main driver. All control variables in (7.2) were highly insignificant apart from $E^i[\mu]$, where β_3 was estimated to equal 0.41(0.10). Participants with higher μ estimates also estimated higher differences in conditional μ^m estimates.

7.3. Other findings from the main experiment

I find that the majority of our sample, 96 per cent, understand the incentive structures underlying the redistribution setting in the experiment: a necessary premise for my model of beliefs to be relevant.

Both egalitarians and libertarians expect worker participants to cast an incentivized vote for no redistribution, and finding no redistribution fair to report more coin flips. Considering the difference within one participant's reported expectations, libertarian participants expect the reporting difference to be 1.6, whereas a libertarian participant expects the difference to be 1. The actual difference in reported coin flips for the worker sample of $N = 18$ is 1.3.

Participants seem to have comprehended the basic premise of incentives underlying the vote for redistribution; 96 per cent answered the comprehension check correctly. I found that men reported significantly more coin flips than women, 7.2 for males versus 5.7 for females. Women also expect others to report significantly fewer coin flips than men. Setting the variable D^{Female} equal to 1 for females and 0 for males, I estimate the regression as follows.

$$E^i[\mu] = \beta_0 + \beta_1 D^{Female} + \epsilon_i \tag{7.3}$$

The estimate of β_1 is $-0.47(0.01)$; women expect others to report 0.5 coin flips fewer than men. Finally, I find projection bias in the reported number of coin flips. We regress participants' reported number of coin flips CR on the expected reported number of coin flips among other participants as follows.

$$E^i[\mu] = \beta_0 + \beta_1 CR + \epsilon_i \tag{7.4}$$

The estimate of β_1 is $0.18(0.0)$; for each additional coin flip, the expected reported number of coin flips among other participants increases by 0.18 flips. Finally, the

participants' risk or political preferences did not correlate with their fairness views, their reported expectations of other participants' coin flips or their own reported coin flips.

7.4. MTurk test

On August 24, 2015, an MTurk pilot was conducted. The full experimental instructions for both pilots can be found in the online pre-analysis plan (Schøyen, 2017). The main goal was to show that elicitation of selfish types was orthogonal to the fairness preference; i.e., egalitarians and libertarians to an equal degree found lying about coin flips an immoral act. This is equivalent to Assumption 1. B of the theoretical model and crucial for the design to measure the intentions behind the redistribution choice. The study was conducted on a sample of 99 American MTurk respondents. Eighty-eight of the 99 respondents reported that they thought it was immoral to report the wrong value on a coin flip in a situation where reporting another value will result in a monetary pay-off. Furthermore, there was zero correlation between attitudes towards misreporting and third-party preferences for redistribution in a setting similar to the main experiment.

Participants also reported that they expected participants voting for no redistribution to report significantly more coin flips and be less likely to donate money to charity in a hypothetical scenario similar to the main experiment. Applying the coin flip measure as a measure of selfishness, the estimate of $E^m[\mu_{nr} - \mu_r]$ was 0.60(0.05), i.e., non-redistributors were expected to report about 50 per cent more favourable coin tosses. Using the number of people in every 100 people who are likely to donate an amount of 5 USD to the non-governmental organization "Save the children" as a measure of selfishness, the estimate of $E^m[\mu_{nr} - \mu_r]$ was 21(0.00); i.e., participants expected 21 fewer people in every 100 no-redistributors to donate to charity.

8. Appendix 4: Experiment instructions

In the following, the experiment instructions are described in order of appearance for each treatment. Before entering the experiment, the subjects in all sessions and treatments were informed that their response is completely anonymous. Before every session, standard introductory information was presented and after every session, information about age, gender and which political party the subject would vote for at a national election was elicited. These questions were similar for all sessions and consequently it is described at the end.

Note that the workers with no incentives over outcomes are referred to as "third party" workers in both experimental instructions and in the headlines of the experimental instructions.

8.1. S.1 Work phase. Participants assigned as workers

Experiment introduction is identical for all participants and described at the end.

8.1.1. S.1 T.1 Selfish action

At your desk, a five kroner coin should be available. When you entered the experiment, the instructor should have shown you how to flip a coin and how to spin the coin. You are now asked to flip or spin the coin 10 times; if your coin lands with the inscription "KONGERIKET NORGE " facing up, you will receive 4 kroner. If your coin lands with the side with the inscription "5 KRONER" facing up, you receive nothing.

How many times did the coin land on the side with the inscription "KONGERIKET NORGE " facing up?

8.1.2. S.1 T.2 Work section

Copy the following text. [See Figure 1.] You are given 15 minutes. If the number of correct words you copy is above the median number copied words in your experimental session, you will be paid 100 kroner. If you copy the median number of words or below you will be paid nothing. The median is a type of mean measure; the "median" is the "middle" value in the list of numbers. If you are not sure about the meaning of "median", please raise you hand and we'll come and assist you. You must finish at least 100 words to complete the task and be eligible for payment and further participation in this work round. The text you write will be matched with the original text, and its similarity will determine how many correct words will be registered.

Words are shown as in Figure.1. The participants have 15 minutes to finish the task. They see a text to copy and can press a menu button for more text.

Mr. Phileas Fogg lived, in 1872, at No. 7, Saville Row, Burlington Gardens, the house in which Sheridan died in 1814. He was one of the most noticeable members of the Reform Club, though he seemed always to avoid attracting attention; an enigmatical personage, about whom little was known, except that he was a polished man of the world. People said that he resembled Byron—at least that his head was Byronic; but he was a bearded, tranquil Byron, who might live on a thousand years without growing old.

Figure 6: Part of the text the participants will copy.

The workers completing the copying task and are assigned their payment.

8.1.3. S.1 T.3 Information about income

You produced over the median number of correctly copied words and are assigned a payment of 100 kroner.

or

You produced the median number of correctly copied words or below, and are assigned a payment of 0 kroner.

We have now assigned either 100 or 0 kroner to all participants in the session who completed the task, in line with the payment schedule we presented before the production phase. As a result, half of the participants that completed the task have been assigned an income of 100 kroner and half of the participants have been assigned an income of 0 kroner as payment for copying text.

8.1.4. S.1 T.4 Vote over relative income

There will now be a vote amongst you and the other members in your experimental session who completed the task. The vote will be on whether or not the income from the experimental session should be equally redistributed amongst the workers within the experimental session. If the redistribution option is chosen, everyone receives 50 kroner. If the no-redistribution option is chosen, the participants producing over median keep the income they were previously assigned. The option that receives the most votes will be implemented. In the unlikely event of a draw an option will be chosen at random.

Before you vote you are asked to answer two questions.

Question.1. Not casting a vote, just asking hypothetically, what do you think is the fairest way to pay the participants of this experiment?
☐ I think the fairest option is the redistribution option.

☐ I think the fairest option is the no-redistribution option.

Question.2. If the redistribution option is chosen everyone receives 50 kroner. If the no-redistribution option is chosen the workers who copied above median number of words keep their earnings of 100 kroner and the other workers receive 0 kroner. To make sure that you have understood the instructions, please tick off which one of the two alternatives below is correct:

☐ The workers who copied above median number of words will make more money if the no-redistribution option is implemented. The workers with median or below number of words will make more money if the redistribution option is implemented.
☐I believe the workers with the median number of words or below words will make more money if the no-redistribution option is implemented. The workers with above median number of words will make more money if the redistribution option is chosen.
You will now cast your vote. Your vote together with the other votes in your group will count in which redistribution option is implemented.

☐ I vote for the redistribution option.

☐ I vote for the no-redistribution option.

Vote is cast and the result of the vote is shown to the participants.

The redistribution option was chosen; *or* The no-redistribution option was chosen.

8.1.5. S.1.T.7 Work phase 2

You now enter the second work phase.

Copy the following text. [See Figure] You are given 15 minutes. If the number of words you correctly copy is above median you will be paid 100 kroner. If you copy the median number of words or below you will be paid nothing. You must finish at least 100 words to complete the task and be eligible for payment.

You produced over the median number of correctly copied words and are assigned a payment of 100 kroner.
or
You produced the median number of correctly copied words or below and are assigned a payment of 0 kroner.

In this round, your payment will be determined by participants who act as third party spectators and who do not have any personal stake in the decision.

Vote casted by third party participants.

If the redistribution option was chosen:

The redistribution option was chosen; you and everyone else will be paid 50 kroner.

If the no-redistribution option was chosen: The no-redistribution option was chosen. You produced the median or under and are consequently paid 0 kroner.

or The no-redistribution option was chosen. You produced over or the median and are consequently paid 100 kroner.

Half of the participants received an income of 100 kroner while the other half receive 0 kroner.

S.1 End

8.2. S.2 Third party

Experiment introduction is identical for all participants and described in the end.

Note that the third parties are in the same session as the workers.

8.2.1. S.1 T.1 Selfish action

At you desk, a five kroner coin should be available. When you entered the experiment, the instructor should have shown you how to flip a coin and how to spin the coin. You are now asked to flip or spin the coin 10 times; if your coin lands with the inscription "KONGERIKET NORGE " facing up, you will receive 4 kroner. If your coin lands with the side with the inscription "5 KRONER" facing up, you receive

nothing.

How many times did the coin land on the side with the inscription "KONGERIKET NORGE " facing up?
times.
All third party participants wait while the workers work.

8.2.2. S.1 T.5 Third parties vote over redistribution

The participants assigned the role as a worker have completed a work task. The work task consisted of copying words from a book for 15 minutes. The participants that completed above the mean number of words, were assigned a payment of 100 kroner for their work. The ones producing the mean or below, were assigned nothing for their work. The median is a type of average measure; the "median" is the "middle" value in the list of numbers. If you are not sure about the meaning of "median", please raise you hand and we'll come and assist you. The participants were aware of this income assignment mechanism before they carried out the task.

There will now be a vote amongst you and the other experimental participants assigned the role of decision maker on whether or not the income of the workers should be equally redistributed amongst the workers within the experimental session.

The option that gets the most votes will be chosen. In the unlikely event of a draw, an option will be chosen at random. If the redistribution option is chosen everyone receives 50 kroner, if redistribution is not chosen the participants keep the

income they were previously assigned.

Before you vote you are asked to answer two questions.

Question.1. Not casting a vote; just asking hypothetically what do you think is the fairest way to pay the participants of this experiment?
☐ I think the fairest option is the redistribution option.

☐ I think the fairest option is the no-redistribution option.

Question.2. If the redistribution option is chosen everyone receives 50 kroner. If the no-redistribution option is chosen the workers who copied above median number of words keep their earnings of 100 kroner and the other workers receive 0 kroner. To make sure that you have understood the instructions, please tick off which one of the two alternatives below is correct:

☐ The workers who copied above median number of words will make more money if the no-redistribution option is implemented. The with workers with median or below number of words will make more money if the redistribution option is implemented.
☐ I believe the workers with the median number of words or below words will make more money if the no-redistribution option is implemented. The workers with above median number of words will make more money if the redistriution option is chosen.

You now cast the counting vote. Please note that the result of the vote will have no economic consequences for the participants that were chosen not to have the role of worker. The outcome of the vote has no economic consequences for you.

☐ I vote for the redistribution option.

☐ I vote for the no-redistribution option.

The vote is cast and the result of the vote is shown to the participants.

The redistribution option was chosen. *or* The no-redistribution option was chosen.

Age, gender, risk- and political preferences are elicited. These questions are similar for all participants and described at the end.

Thank you for your participation!

8.3. S3. Prediction session

S.3 In the prediction phase, a new group of experimental subjects referred to as predictors will make incentivized guesses of the overall portions of workers and third parties that made which choices. They will also guess what the correlation between the choices made in S.1. and S.2. The predictors make their guesses in a separate sessions after the work phase

8.3.1. S.3 T.1 Prediction of worker behavior

Experiment overview. You are to guess what portion of the participants acted in what way in a previous experimental session previously held at this school. The experiment and all the relevant details will be described to you. The data from the experiment has been recorded and your payment will be higher the more precise your estimate the data. The precision of your estimate is measured by a quadratic function. You are paid increasingly more the more precise you are, and if you guess the exact value you will be paid 20 kroner. If it is off by more than 15 percent, you will not be paid for the estimate. For details on the scoring rule press here.

Link to exact description of scoring rule.

After the experiment has been described to you, you will be shown the exact menus and choices offered at the first session.

As mentioned, an experimental session was held earlier today *or* yesterday in this room with participants from your school. In the experiment, participants assigned the role of workers did two different tasks.

First the participants were asked to flip or spin a Norwegian 5 kroner coin 10 times. Before flipping the coin they were told that each time the coin landed with the head side up they would receive 4 kroner. No one else than the participants themselves would observe the result of the coin flip and there would be no way after flipping the coin to check how it landed. The coin given has equal probability of landing on each side so that if all participants reported truthfully, half of them would rapport the coin landing on"KONGERIKET NORGE ".

After completing this task, the participants assigned the role of workers did a second task; a work task. The work task consisted of copying words from the book "Around the world in 80 days " by Jules Verne for 15 minutes. The participants that completed above the median number of words received a payment of 100 kroner for their work, the other half, the ones producing the median number of words or below , received nothing for their work. The median is a type of average measure; the "median" is the "middle" value in the list of numbers. If you are not sure about the meaning of "median", please raise you hand and we'll come and assist you.

The workers were aware of this payment mechanism before they did the task. The participants could then vote for whether or not the final payment should be equally redistributed among the participants. Before voting they answered two question. Firstly they were asked what option they thought was the fairest option. Secondly the participant were asked what option they thought which group would make more money on if implemented. There were then two options they could vote for:
□ The redistribution option; every worker that had completed the task in the session received 50 kroner.
□ The no-redistribution option; every worker received the payment he or she was assingned to receive.
The option that received the most votes was implemented and in the very unlikely event of a draw, an option was chosen at random.

You will now be shown screen shots of exactly the same menus as seen by the workers. These are shown to make sure you understand how the previous experimental session was conducted. *Screen shots of the menus of the worker saw with a clear*

label saying this is what the workers saw are shown to the predictors.

Prediction phase:
You are now to predict how the students acted.

Out of 100 experimental subjects assigned the role of workers, how many do you think chose the redistribution option?

What number of "KONGERIKET NORGE " do you think was the average reported among the worker participants? (Please note that they were aware that "KONGERIKET NORGE " was the side giving a payment of 4 kroner. Please also note that if everyone reports honestly, the expected average will be 5.)

What number of "KONGERIKET NORGE " do you think was the average reported among the worker participants that voted for redistribution?

What number of "KONGERIKET NORGE " do you think was the average reported among the worker participants that voted for no redistribution?

8.3.2. S.3 T.2 Prediction of third party behavior

There was a second round of work for the worker participants. In this round of work, the decision of whether to redistribute or not was made by a vote by other experimental subject assigned the role of third-party decision makers. The workers completed their task knowing that other participants would decide whether their

earnings should be redistributed. The workers were then paid in accordance to the majority choice of the experimental subjects assigned the role of decision makers. These subjects voted for the redistribution option they found the most fair. Their payment was simply the show up fee and they had no material interest to vote for any of the alternatives. Similarly to the workers, they were asked two questions before voting. Firstly they were asked what option they thought was the fairest option. The participant were then asked which group would make more money on what redistribution option being implemented. You will be shown the exact menus and choices they were given.

The predictors are shown screen shots of the same menus as seen by the third parties .

Out of 100 experimental subjects assigned the role of third-party decision maker, how many do you think chose the redistribution option?

They were also given a coin to flip a five kroner coin ten times and were to rapport how many times this coin landed on the side with the inscription "KONGERIKET NORGE ". They were told they would be paid 4 kroner for each time they reported flipping the coin on the coin landed on the side with the inscription "KONGERIKET NORGE ".

What number of "KONGERIKET NORGE " do you think was the average reported among the third-party decision makers that voted for redistribution ?

What number of "KONGERIKET NORGE " do you think was the average reported

among the third-party decision makers that voted for no redistribution ?

You are now asked some questions about your background and preferences.

8.3.3. S.3 T.3 Predictors vote over redistributive preferences

Another experimental session will be held after this. In this session the participants assigned the role of worker will complete a work task exactly equal to the work phase in the one previously described. The work task will consist of copying words from a book for 15 minutes. The one half of the participants that complete above the median number of words receive a payment of 100 kroner for their work, the other half, the ones producing the median or below, receive nothing for their work. The participants will be aware of this payment mechanism, but not the redistribution options, before they do the task.

There will now be a vote on whether or not the income from the experimental session should be redistributed amongst the workers within this future experimental session. The option which receives the most votes will be chosen, in the event of a draw an option will be chosen at random. If the redistribution option is chosen every participant that completed the task receives 50 kroner, if redistribution is not chosen the participants keep the income they previously assingned.

Before you vote you are asked to answer two questions.

Question.1. Not casting a vote; just asking hypothetically, what do you think is the fairest way to pay the participants of this experiment?
☐ I think the fairest option is the redistribution option.

☐ I think the fairest option is the no-redistribution option.

Question.2. If the redistribution option is chosen everyone receives 50 kroner. If the no-redistribution option is chosen the workers who copied above median number of words keep their earnings of 100 kroner and the other workers receive 0 kroner. To make sure that you have understood the instructions, please tick off which one of the two alternatives below is correct:

☐ The workers who copied above median number of words will make more money if the no-redistribution option is implemented. The workers with median or below number of words will make more money if the redistribution option is implemented.
☐I believe the workers with the median number of words or below words will make more money if the no-redistribution option is implemented. The workers with above median number of words will make more money if the redistriution option is chosen.
You will now cast your vote; this vote together with the other votes in your group will count in which redistribution option is implemented.

Please note that the result of the vote will only have economic consequences for the participants that were chosen to have the role of worker. The vote has no economic consequence for you.
☐ I vote for the redistribution option.

☐ I vote for the no-redistribution option.

8.3.4. S.3 T.4 Elicitation of selfishness

At you desk, a five kroner coin should be available. When you entered the experiment, the instructor should have shown you how to flip a coin and how to spin the coin. You are now asked to flip or spin the coin 10 times; if your coin lands with the inscription "KONGERIKET NORGE " facing up, you will receive 4 kroner. If your coin lands with the side with the inscription "5 KRONER" facing up, you receive nothing.

How many times did the coin land on the side with the inscription "KONGERIKET NORGE " facing up?

.....

times.

8.3.5. S.3 T.5 Predictors are paid according to accuracy

You will be called up by experimental id number when you can collect your earnings.

Your total payment is kr XXX

S.3 End

8.4. Introduction, elicitation of age, gender political preferences and risk preferences

After all of the sessions, participants are elicited for political preferences, risk preferences and gender.

8.4.1. Introduction

General overview text:

Welcome as a participant in this experiment. Please read the instructions carefully. In economic experiments such as this one, the experimenter will never lie to or manipulate, the experimental subjects such as yourself. Consequently all the information you will be given is accurate and not misleading in any way. You will be given all the information you need. Please do not communicate with the other participants during the experiment. If you have any questions, please raise your hand and ask us. All the information we will use from you is the one you provide; we are not recording you in any other way. The information will be treated anonymously.

8.4.2. Political preferences

If it was the general elections, (Stortingsvalg), in Norway today what party would you vote for?
If you do not have the right to vote in Norway, please answer the question as if you did have a right to vote.

Options:

☐ Labour Party [Arbeiderpartiet]

☐ Conservative Party [Høyre]

☐ Progress Party [Fremskrittspartiet]

☐ Centre Party [Senterpartiet]

☐ Liberal Party [Venstre]

☐ Socialist Left Party [Sosialistisk Venstreparti]

☐ Green Party [Miljøpartiet De Grønne]

☐ I would not vote

☐ I would not vote for any of the alternatives given here

8.4.3. Gender

What is your gender?

☐ Male

☐ Female

☐ Transgender

8.4.4. Age

What year were you born?

[Enter year]

8.4.5. Risk preference elicitation

How do you see yourself: are you generally a person who is fully prepared to take risks or do you try to avoid taking risks? Please tick a box on the scale, were the value 0 means: 'not at all willing to take risks' and the value 10 means: 'very willing to take risks'

[*Enter number between 1 and 10*]

8.4.6. Precise description of the scoring rule.

Simply put the quadratic scoring rules pays you progressively more the closer you are to the true value. By progressively is meant your payment increases quadratically the closer your estimate is to the true estimate value.

The quadratic scoring rule is

$$\pi = A - K(\text{estimate - true value})^2$$

. Where π is how much you are paid. For all estimates A = 20 kroner. For estimates in the 0-100 range $K = \frac{1}{11.25}$ and for estimates in the 0-10 range $K = \frac{1}{1.125}$. If the value of π is negative you will simply be paid nothing for your estimate.

9. References

Acemoglu, Daron and Alexander Wolitzky. 2014. "Cycles of conflict: An economic model." *The American Economic Review* 104(4):1350–1367.

Alicke, Mark D, David A Dunning and Joachim Krueger. 2005. *The self in social judgment*. Psychology Press.

Binder, Sarah. 2015. "The dysfunctional congress." *Annual Review of Political Science* 18:85–101.

Blanco, Mariana, Dirk Engelmann, Alexander K Koch and Hans-Theo Normann. 2014. "Preferences and beliefs in a sequential social dilemma: a within-subjects analysis." *Games and Economic Behavior* 87:122–135.

Bushong, Benjamin and Tristan Gagnon-Bartsch. 2016. "Learning with Misattribution of Reference Dependence." *Working Paper* .

Cappelen, Alexander W, Astri Drange Hole, Erik Ø Sørensen and Bertil Tungodden. 2007. "The pluralism of fairness ideals: An experimental approach." *American Economic Review* 97(3):818–827.

Cohn, Alain, Ernst Fehr and Michel André Maréchal. 2014. "Business culture and dishonesty in the banking industry." *Nature* 516(7529):86–89.

Gagnon-Bartsch, Tristan. 2017. "Taste Projection in Models of Social Learning." *Working Paper* .

Graham, Jesse, Brian A Nosek and Jonathan Haidt. 2012. "The moral stereotypes of liberals and conservatives: Exaggeration of differences across the political spectrum." *PloS one* 7(12):e50092.

Haidt, Jonathan. 2007. "The new synthesis in moral psychology." *Science* 316(5827):998–1002.

Haidt, Jonathan. 2012. *The righteous mind: Why good people are divided by politics and religion*. Vintage.

Harsanyi, John C. 1962. "Bargaining in ignorance of the opponent's utility function." *Journal of Conflict Resolution* pp. 29–38.

Madarász, Kristóf. 2015. "Projection equilibrium: Definition and applications to social investment and persuasion." *Working paper* .

Paluck, Elizabeth Levy. 2012. "Interventions Aimed at the Reduction of Prejudice and Conflict." *The Oxford Handbook of Intergroup Conflict* p. 179.

Piketty, Thomas. 1995. "Social mobility and redistributive politics." *The Quarterly Journal of Economics* 110(3):551–584.

Reeder, Glenn D, John B Pryor, Michael JA Wohl and Michael L Griswell. 2005. "On attributing negative motives to others who disagree with our opinions." *Personality and Social Psychology Bulletin* 31(11):1498–1510.

Roemer, John E. 2009. *Equality of opportunity*. Harvard University Press.

Ross, Lee, David Greene and Pamela House. 1977. "The "false consensus effect": An egocentric bias in social perception and attribution processes." *Journal of Experimental Social Psychology* 13(3):279–301.

Rubinstein, Ariel and Yuval Salant. 2016. "Isn't everyone like me?: On the presence of self-similarity in strategic interactions." *Judgment and Decision Making* 11(2):168.

Schøyen, Øivind. 2017. "PRE-ANALYSIS PLAN: Attributing intentions under projection bias." *American Economic Association RCT Registry* (2310).

Chapter 3

Co-authored by Xianwen Chen, The Norwegian Institute for Nature Research

Paternalist motivation: An experimental test

Abstract

Is people's willingness to implement their fairness views on a group dependent on how many in the group share their view? We designed a new lab experiment to answer this question. Spectator participants were asked how many other people they believe share their view of whether or not it is fair to redistribute income in a work task where the output is determined by luck and effort. They were then given the option to pay two cents to implement the distribution they found fair, upon a pair of worker participants who had completed the work task. We find that willingness to pay to implement the distribution the spectator participants found fair is completely uncorrelated with their beliefs about how many share their view among the affected worker participants. Furthermore, although spectator participants systematically overestimate how many share their fairness view, being informed about the true number does not affect their decision to implement the distribution they found fair. The results suggest that people have paternalist motivation to implement their views of the world: they are motivated to implement their own fairness view, regardless of whether their view is at odds with that of those who are affected. We discuss how paternalist motivation has implications for delegating collective choice in groups where fairness views vary, and how our finding contributes to our understanding of why political polls affect voting.

1. Introduction

People often exert effort to have their views of the world implemented, beyond what can be rationalized by their material or private interests. Examples include participation in civil society, voluntary political activities, revolutions and voting; activities motivated by changing the world for fellow humans.

How does willingness to implement one's views relate to the preferences of those affected? It could be motivated by a belief that the affected share one's preferences for how the world should be, non-paternalist motivation. Alternatively, the motivation could be unrelated to what the affected prefer, paternalist motivation. Non-paternalist motivation can be rationalized by a belief that the affected would fare better if the world was in accordance with their own preferences, or a belief that people have a right to live in a world in accordance with their views. Paternalist motivation could arise from a belief that the affected will fare better under certain states of the world, independent of their preferences, or a belief that the world ought to be in accordance with some principle. Paternalist motivation is negatively defined as *not* related to the preferences of the affected. Hence, paternalist motivation could also arise from much less sophisticated motivations, i.e., people might have an innate tendency to seek to align the world to their own views.

Motivation to implement states upon others is important for understanding whether decision makers seek to make socially desirable decisions for groups with mixed views, or merely implement their own view. If decision makers have a purely non-paternalist motivation, civic engagement and delegation in bureaucracies function smoothly as an aggregation of preferences for groups. In contrast, given purely paternalist motivation, decision makers will only implement the state of the world they find fair. This will, in many instances, have socially undesirable consequences for those af-

fected. Hence, paternalist motivation can inform policy makers seeking to design institutions that facilitate mixed views. An example of an applied question relating to paternalist motivation is to what degree teachers can be delegated responsibility for curriculums covering politically debated subjects.

To investigate paternalist motivation, we study how willingness to pay to implement redistribution is related to whether the affected share the decision maker's fairness view. Previous studies have shown that people hold mutually excluding views of whether or not income should be redistributed in settings where both luck and effort determine output (Cappelen et al., 2007). Some find it fair to redistribute earnings, others do not. For a person with non-paternalistic motivation, a willingness to impose redistribution is given by how many in an affected group she believes find redistribution fair. However, if she has a purely paternalistic motivation for implementing redistribution, there is no reason to expect that the share of individuals sharing her view will affect her decision.

In our experiment, we randomly assign participants to one of two roles; participants are either assigned the spectator role of making decisions for which they have no private incentives, or, the worker role of completing tasks for payment. We then study if spectators have a willingness to pay to have the workers' payment distributed in the way the spectators find fair. If spectators have a non-paternalist motivation, we should expect to observe the following two patterns in the experiment: *(I)*. Information about the true number of workers sharing the spectator's fairness view should be relevant for spectators with non-paternalist motivation in deciding whether or not to implement a distribution choice. More concretely, if a spectator receives information that fewer workers than she expects share her view, then this information should reduce her non-paternalist motivation to implement her view. Furthermore, assuming that beliefs and preferences are uncorrelated, we should also expect: *(II)*.

The willingness to pay to impose the spectator's perceived fair distribution should be higher for spectators with higher estimates of how many people share her fairness view among the workers.

Our experiment design is as follows. We first ask what spectators perceive to be a fair way to distribute income in a hypothetical scenario. In the scenario, a pair of workers are assigned the task of identifying which letter was next to which number on a list. Each worker is assigned a separate randomly drawn price between one and 10 cents per correctly identified number–letter combination. The distribution alternatives are either full redistribution, workers earn a payment equal to their combined earnings divided by two, or no redistribution such that workers are paid their earnings. Spectators are then told that some workers have completed this task, and asked how many of the workers they believe share their view of what is the fair distribution; we refer to this as the spectator's prevalence estimate of fairness views among the workers. They are then given a choice to pay two cents of their participation payment to implement the option they report to be the fair option for a pair of workers who had completed a task identical to the one in the hypothetical scenario. If they do not pay, the distribution option they find to be the least fair is implemented upon the workers. In one treatment session, spectators are informed about the true prevalence of their fairness view among the workers. The information is given after the spectators reported their expected prevalence, but before deciding whether or not to pay to implement the distribution alternative they find fair. This is done to see if the number of workers sharing the spectator's fairness view is relevant information to the spectator's decision of whether to implement the redistribution she finds fair.

This paper provides three findings. *First*, we find that two-thirds of spectators are willing to pay to have their fairness view implemented upon a pair of workers; the

share implementing is similar among spectators of both fairness views. *Second,* we find spectators' willingness to implement what they consider to be a fair distribution completely uncorrelated with their perception of how many share their fairness view among the workers. *Third,* and the main finding, is that there is no effect of information about the true prevalence of fairness views on the probability that a spectator paid to implement the redistribution she found fair.[1] Although participants overestimate the prevalence of their fairness view among the workers, there is no evidence that being informed about the true prevalence affects their decision of whether or not to implement. This is true for spectators of both fairness views; being informed about the true prevalence had no effect both for spectators who found their view to be shared by a majority, or a minority, of the workers. Furthermore, being informed about the true prevalence had no effect for spectators with particularly high prevalence estimates. The absence of any effect of being informed about the true prevalence of fairness views, combined with the overestimation of prevalence, provides causal evidence suggesting that spectators lack non-paternalist motivation; the large share of spectators implementing is evidence that spectators have paternalist motivation. The findings suggest that the motivation for implementing fairness views reflects a notion of how redistribution *ought* to be, rather than a concern for what those affected find to be fair. Our analysis is guided by a simple theoretical framework; we clarify the assumptions required to interpret our findings as a rejection of the presence of non-paternalist motivation among spectators.

The paper relates to three strands of literature: paternalism, the motivation for voting, fairness views and redistribution. Our definition of paternalist motivation

[1]The second and third findings were surprising to us and go against our pre-specified hypothesis (Chen and Schøyen, 2017).

strays subtly from the classic notion of paternalism as acting to avoid others from falling in harm's way against their own will (Coons and Weber, 2013). It contributes to the literature on paternalism (Coons and Weber, 2013; Thaler and Sunstein, 2003) by considering mutually excluding fairness views and abstracting from the question of harm by introducing the concept of paternalist motivation. This is done to discuss the motivation for implementing states of the world adhering to mutually excluding ideas of what constitutes fairness. When there are mutually excluding alternatives that might be conceived as fair, such as whether or not to redistribute income, it is ambiguous if any party has indeed suffered harm. The question of harm is then contingent on the eye of the beholder; one woman's perception of fairness is another's perceived unfairness. The role of multiple conflicting views of what is fair is a fundamental limit to the desirability of liberal paternalism (Thaler and Sunstein, 2003). Our results and concept of paternalist motivation can be extended to a framework to test and formalize libertarian words of caution; this is of applied interest in societies with strong political polarization, or that are multicultural.

The paper also relates to the literature on fairness views and redistribution. Previous studies by Durante, Putterman and Van der Weele (2014) and Buser, Putterman and van der Weele (2016) investigate how participants' willingness to pay to implement their preferred tax rate upon others relates to self-interest, insurance motives, social concerns and gender. This study establishes that the prevalence of the spectators' fairness view among those affected is not relevant for spectators when deciding whether to implement fairness views upon groups. Furthermore, the study is novel in focusing on the motivation in a setting where the spectator has not been a participant in a setting equal or similar to any of the affected workers.

Finally, the paper contributes to the literature on the motivation for voting: a costly type of implementation of social preferences. Paternalist motivation is important

for understanding why people prefer to vote for political parties with good opinion polls (Rothschild and Malhotra, 2014). This effect could be caused by opinion polls being a social signal on what is the right way to vote; the effect could also be a sign of non-paternalist motivation, i.e., the electorate is also the affected group and their preferences are relevant for people's non-paternalist motivation to vote. The literature on the effect of opinion polls on voting focuses on different aspects of the expressive utility of votes, including normative social influence, informational social influence, resolving cognitive dissonance by switching to the side they infer is going to win and the tendency for conformity. Our study differs from previous studies as people have already expressed their fairness preference before being informed about the true prevalence of their view. Hence, we focus on the component of voting motivation driven by a wish to implement outcomes out of fairness concerns. As we find no evidence of non-paternalist considerations, the finding is consistent with the effect opinion polls have on voting that stems from voters changing their minds about the proper way to vote through social signals of what is the "correct" way to vote.

The paper proceeds as follows. First, we make clear our definitions of paternalist and non-paternalist motivation in a brief theory section. We then describe the experiment design and results, before the paper concludes.

2. A theory of spectator motivation for implementing a state upon a group

We consider the utility of a rational decision maker i implementing a state s upon a group $\boldsymbol{J}$ consisting of $N \geq 2$ individuals. We develop definitions of non-paternalist and paternalist motivation. We allow for i to have both types of motivation, and consider how her utility of implementing s changes with her type of motivation and her estimate of how many in $\boldsymbol{J}$ share her fairness view.

The world can be in either the state s or its complement state s^c; we denote a generic state of the world w such that $w \in \{s, s^c\}$. We assume there is a monetary cost difference between s and s^c for i; the difference between incurring the monetary cost in s and s^c decreases i's utility by some constant $c_i > 0$. Whether the state of the world is s or s^c changes outcomes for all members of the group $\boldsymbol{J}$; it affects i only through her other-regarding preferences and by her cost of implementing it.
We first consider non-paternalist motivation. We denote an arbitrary individual in the affected group $\boldsymbol{J}$ as j. Defining a person i's non-paternalist utility of imposing some state w^c upon an individual j who finds its complement state w to be the fair state, is defined as $v_i^w(w^c)$. We define non-paternalist motivation as follows.

Non-paternalistic motivation Assume a person j finds state w to be more fair than its complement state w^c. Then a person i with non-paternalist motivation will receive strictly higher non-paternalist utility from imposing w upon j than imposing w^c, $v_i^w(w) > v_i^w(w^c)$.

A spectator i's difference in non-paternalist utility between implementing w or w^c for a person j finding either w or w^c more fair is then:

$$\Delta v_i^w \equiv v_i^w(w) - v_i^w(w^c),\ \Delta v_i^{w^c} \equiv v_i^{w^c}(w^c) - v_i^{w^c}(w). \tag{1}$$

We define that the decision maker i believes that a portion, $\alpha_i(s)$, of individuals in $\boldsymbol{J}$ prefer the state s.

$$\alpha_i(s) \equiv E_i \left[\frac{\text{Number of Individiduals in } \boldsymbol{J} \text{ preferring } s}{N} \right] \tag{2}$$

The expected non-paternalist utility difference for i between imposing s on the group $\boldsymbol{J}$ is then as follows.[2]

$$N\left(\alpha_i(s)\Delta v_i^s - (1-\alpha_i(s))\Delta v_i^{s^c}\right) \tag{3}$$

We now define paternalist motivation.

Paternalistic motivation A person i with paternalistic motivation has a positive utility differential between imposing state s and complement state s^c upon a group of individuals, $\boldsymbol{J}$. $\kappa_i > 0$, if and only if i finds s to be fairer than its s^c.

Assume that i is aware that j finds s^c to be the fair state. The utility for j is not relevant to the paternalist motivation of i; i will have some utility $\kappa_i > 0$ of implementing s upon j regardless of the views of j.[3]

We now consider i's decision to implement s upon $\boldsymbol{J}$ under both paternalist and non-paternalist motivation. Assume that the decision maker finds the state s to be the fair state of the world; all arguments hold, ceteris paribus, if she finds s^c to be

[2]Note the relationship between utilitarianism and non-paternalist motivation. Suppose the following two relations hold. *(I)*. An individual j affected by state s has a higher utility of living under the state she finds fair, and i is aware of this. *(II)*. i is a utilitarian who will receive some utility from increasing the overall utility of individual j. If *(I)* and *(II)* hold, it follows that she will have some utility from imposing the state that j finds fair rather than the state that j finds less fair; this coincides with the definition of non-paternalist motivation.

[3] Note that paternalistic motivation can arise from what we refer to as a Kantian moral motivation. Suppose that i finds the state s to be inherently morally right; the state has inherent qualities that exist independent of the sentiments of or utility for individuals in $\boldsymbol{J}$ or i herself. See for instance Rosati (2006) for a discussion of Kantian moral motivation.

the fair state. The expected utility for i to implement s upon $\boldsymbol{J}$, when the alternative to implementing is s^c, is then as follows.

$$E_i[\Delta u_i^s] = N\left(\alpha_i(s)\Delta v_i^s - (1-\alpha_i(s))\Delta v_i^{s^c}\right) + \kappa_i - c_i \tag{4}$$

Note that the expected utility of implementing the state varies between decision makers for both the different components of non-paternalist utility $\left(\alpha_i(s)\Delta v_i^s - (1-\alpha_i(s))\Delta v_i^{s^c}\right)$, the size of the group N, the paternalist utility κ_i and the utility loss of implementing c_i. The decision maker will implement state s whenever (utilf2) is positive. By the definition of Δv_i^w in (harkonen) and the definition of non-paternalist motivation, it holds that $\Delta v_i^s > 0$ and $\Delta v_i^{s^c} > 0$. From this and (utilf2), it follows directly that decision makers with higher prevalence estimates and more non-paternalist motivation will have a higher utility of implementing s, i.e., $\frac{\partial E_i[\Delta u_i^s]}{\partial \alpha^i(s)} = N[\Delta v_i^s + \Delta v_i^{s^c}] > 0$.
We now briefly discuss the probability that a random decision maker in a group of decision makers $\boldsymbol{I}$ implements s, $[P_i = 1]$. If the preference parameters Δv_i^s and $\Delta v_i^{s^c}$ are drawn from distributions that are independent of the distribution of the belief parameter α_i, the probability that a random decision maker implements s, will be positively correlated with her prevalence estimate $\alpha_i(s)$. Furthermore, if decision makers are given information about the true prevalence of a fairness view among the workers, implying that it is lower than their expected prevalence, it will lower the utility of implementing and weakly decrease the probability of a random spectator implementing; the size of the effect depends on how many decision makers in the group have their utility changed from positive to negative by the information, lowering their $\alpha_i(s)$. Given that all parameter distributions have continuous support for their entire domain and the size of the group of decision makers is infinitely large, information increasing the median $\alpha_i(s)$ in a group will strictly increase the expected

number of spectators in a group of spectators who are implementing.

3. Experimental design

We now present the experiment design. Participants are randomly assigned to one of two roles: worker or spectator. The main focus of the study is on the spectators' choices; the workers complete their task so that the choices made by spectators are about concrete outcomes rather than hypothetical scenarios.

First, workers complete two rounds of work where they are given one minute to identify which number is next to a specified letter on a list of numbers and letters. Each worker draws a unique price between one and 10 cents; the worker earns this price per correctly identified number–letter combination. Before completing the work task, the workers are asked whether they find it fair to redistribute earnings in a scenario of a pair of workers who have completed a task identical to the one they are to perform subsequently. The redistribution alternatives are that either workers are each paid an equal half of their combined earnings, referred to as the redistribution option, or that the workers are paid a wage according to their separate earnings, referred to as the no-redistribution option.

Second, the two spectator sessions, a control session and a treatment session, are held simultaneously. Spectators are randomly allocated to sessions. In both sessions, the spectators are first presented the same hypothetical scenario as the workers, i.e., a pair of workers had completed tasks and have earned a random price per solved task. The spectators are then asked whether they find the redistribution or the no-redistribution option fair. In the control session, the spectators are then given a choice to pay two cents to implement their preferred redistribution upon a pair of workers. The cost of implementing the spectator's view of redistribution, two cents, is deliberately chosen to be low, to permit identification of very "weak"

preferences for implementing. They are informed that if they do not pay two cents to implement their preferred distribution, the distribution they found least fair will be implemented. In the treatment-group spectator session, the spectators are informed about the true prevalence of their redistribution preferences among the workers after giving their prior belief, but before they make the choice of whether to pay two cents to implement it. This is done to investigate whether knowing the true prevalence will affect the probability of a spectator being willing to pay to implement her third-party preference.

A pre-analysis plan, Chen and Schøyen (2017), was posted online before the experiment was conducted. The experiment design session is described in Appendix B, and the experiment instructions in Appendix C.[4]

4. Results

We conducted the experiment on Amazon's online platform for experiments, MTurk, where we recruited participants with IP addresses registered in the United States of America. We first gathered data from 336 workers. Among the workers, 33% found redistribution fair, while 67% found no-redistribution fair. The information treatment consisted of informing spectators of this prevalence.

Table tab: descriptive statistics breaks down the spectators by control or treatment group, by their fairness view and by whether they implemented the distribution they found fair. Spectators overestimate how many share their fairness view; this finding, defined as projection bias (Ross, Greene and House, 1977), is in line with previous

[4] Participants were randomly allocated to one of the two sessions with equal probability until the total sample size of 672 was reached. The sample sizes differ in the control group and the treatment group because of small sample issues with our true randomization algorithm.

studies on projection bias across a wide number of settings (Ross, Greene and House, 1977; Alicke, Dunning and Krueger, 2005).

Result 1: The majority of spectators are willing to pay to implement their fairness view.

About two-thirds of spectators are willing to pay two cents to implement their preferred redistributive option. Fairness views are uncorrelated with willingness to pay to implement. The portion of participants implementing is stable at two-thirds for both fairness views and treatment or control session; this can be seen from the second column labelled "Percentage of spectators" in Table tab: descriptive statistics.

Result 2. Willingness to implement fairness views is uncorrelated with prevalence estimates $\alpha_i(s)$.

Spectators' willingness to implement their fairness view is uncorrelated with their prior prevalence estimate of how many share their view $\alpha_i(s)$. This can be seen from pairwise comparing average prevalence estimates in the far right column of Table tab: descriptive statistics. The average prevalence estimates are similar for participants choosing to implement or not implement. Figure fig: Histogram of AplhaEasy shows the portion of spectators choosing to implement according to whether their prior belief of prevalence was above or below the actual prevalence among the workers, or accurate. Prevalence estimates are on average similar in both the information treatment where the true prevalence was revealed and in the control treatment, which can be seen from Figure fig: Histogram of AplhaEasy.

We regress the probability of a spectator in the control group implementing her preferred redistribution s, $[P_i = 1]$ on her prevalence estimate and on how many

Table 1: Descriptive Statistics

Treatment Group	Fairness view	Number of spectators	Implemented or not	Percentage of spectators	Prevalence estimate
No information	Redistribution	85	Yes	69%	50.37%
			No	31%	53.77%
	No-redistribution	222	Yes	68%	70.23%
			No	32%	69.07%
Information	Redistribution	105	Yes	62%	52.03%
			No	38%	57.83%
	No-redistribution	260	Yes	70%	71.15%
			No	30%	68.81%
Sample size		672			

Note: The first column from the left labelled treatment group divides spectators into categories according to whether they received the information treatment or not. The second column from the left labelled fairness view shows how many in the control or treatment categories found redistribution or no-redistribution fair. The third column labelled number of spectators shows the number of spectators having different fairness views for the different categories. The fourth column labelled implemented or not shows the percentage of spectators implementing their fairness view for treatment, fairness view and whether they implemented. The far right column labelled prevalence estimate shows the average prevalence estimate for participants in each category.

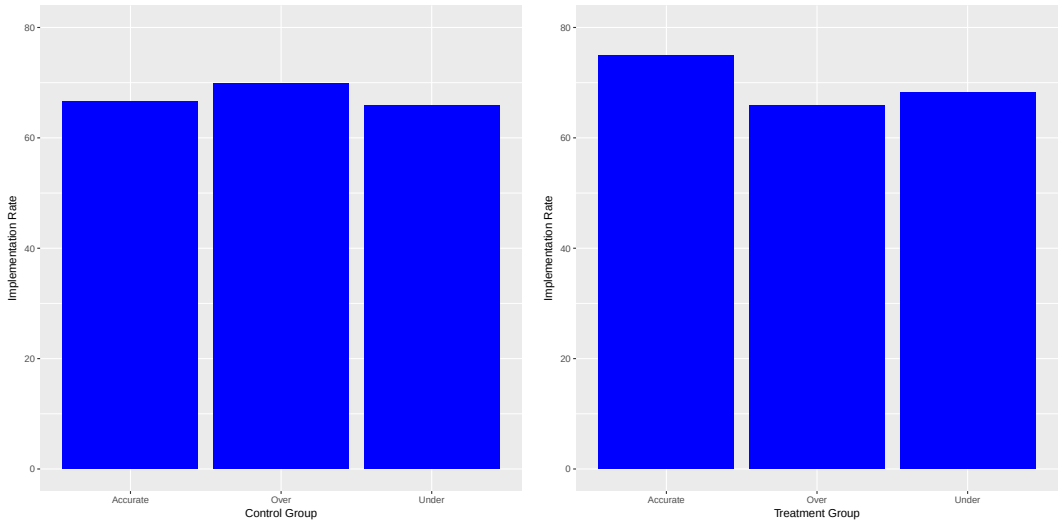


Figure 1: Histogram of percentage of spectators choosing to implement according to their prevalence estimate of the fairness view among workers. The categories reflect whether the spectator estimate of how many shared her fairness view among a hundred workers was over the actual, under the actual or accurate within an interval of plus/minus five workers.

found her preferred form of redistribution s fair $\alpha_i(s)$.

$$[P_i = 1] = \beta_0 + \beta_1 \alpha_i(s) + \epsilon_i \tag{5}$$

Table 2 shows the results from estimating (first): β_1 is estimated to be -0.0002 (0.900) for the control group.[5] The result is robust for the pooled sample and the treatment group data, as shown in Table 2.

Figure fig: Histogram of Aplha shows the distribution of prevalence estimates for subsamples of fairness views and whether participants are in the treatment or control group. The bars show total number of spectators with a prevalence estimate within the bin on the horizontal axes. The bars are divided into spectators choosing to implement or not to implement the distribution they find fair. Comparing those who prefer no-redistribution in the control and treatment groups, the distributions of prevalence projections are similar, so are the percentage of spectators willing to pay to implement. The same holds for those who prefer redistribution in the control and treatment groups. To confirm our graphical observation, we tested for differences in distributions of projected prevalence between spectators choosing to implement or not implement, using a Kolmogorov–Smirnov test for equal distribution (0.9672) for the main sample and all subsamples in Figure fig: Histogram of Aplha. Details about the test and further results can be found in Appendix sec: KS tests. Assuming that the prevalence estimates and paternalist motivation are not negatively correlated,

[5]The probability of no effect given the observed estimate, P-value, are stated in parenthesis throughout the main paper.

Table 2: Estimation of (first): Correlation between willingness to implement fairness view and prevalence estimate of own fairness view, $\alpha_i(s)$.

	Dependent variable: $[P_i = 1]$		
	Pooled sample	Control group	Treatment group
$\alpha_i(s)$	0.0002	-0.0002	0.001
	(0.001)	(0.001)	(0.001)
Constant	0.667^{***}	0.695^{***}	0.644^{***}
	(0.059)	(0.087)	(0.080)
Observations	672	307	365
R^2	0.0001	0.0001	0.001
Adjusted R^2	-0.001	-0.003	-0.002
Residual std. error	0.467 (df = 670)	0.466 (df = 305)	0.469 (df = 363)
F statistic	0.056 (df = 1; 670)	0.016 (df = 1; 305)	0.190 (df = 1; 363)
Note:	*p<0.1; **p<0.05; ***p<0.01		

this finding is supportive of the absence of paternalist motivation.[6]

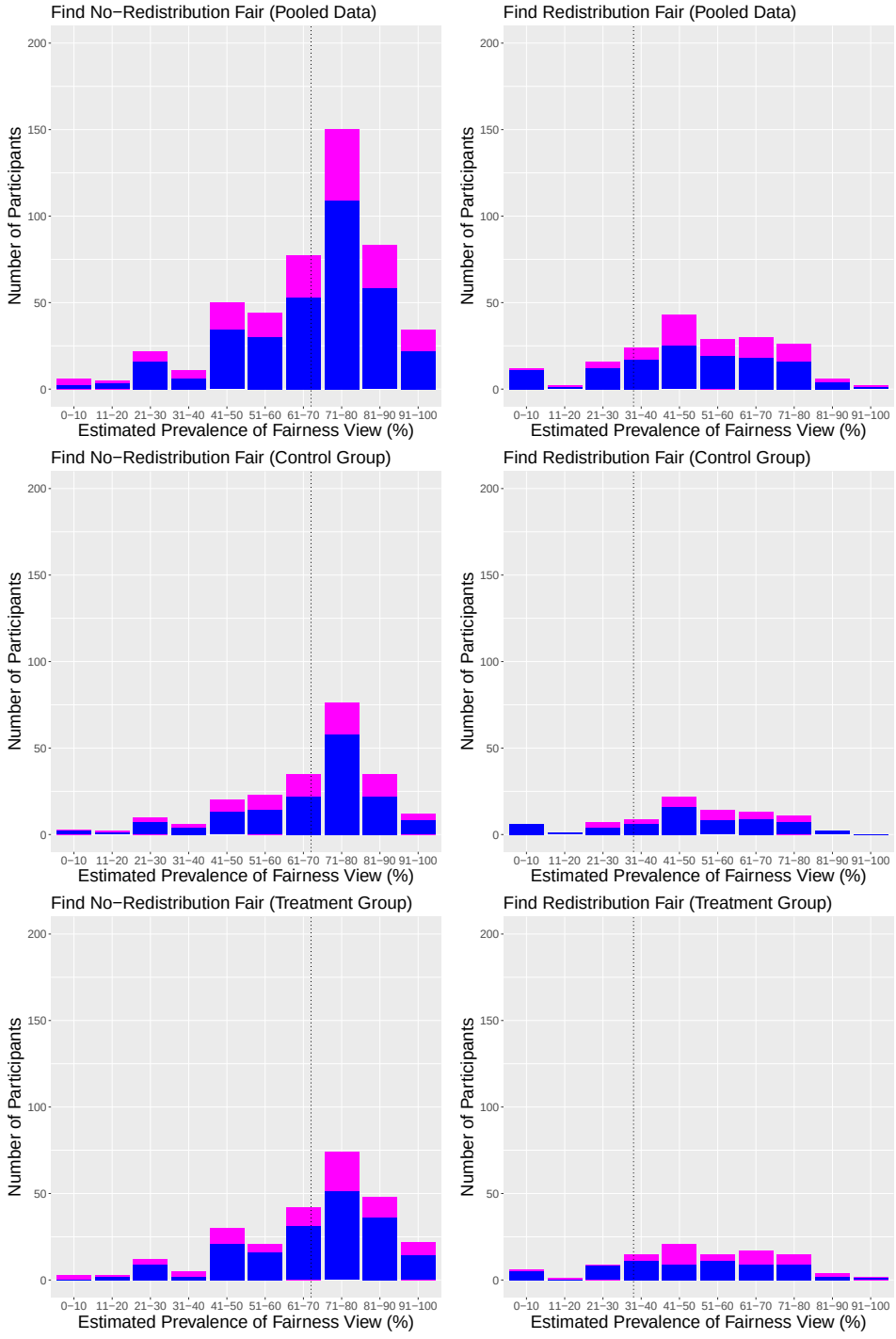
Result 3. There is no causal effect of receiving information about true prevalence $\alpha^W(s)$ on spectators' willingness to implement fairness view.

The third and main finding of our study is that being informed about the true prevalence of their fairness view had no effect on the probability that spectators implemented the redistribution alternative they found fair. The spectators in the treatment session are shown the true prevalence of their redistribution preference among the workers before deciding whether or not to implement. The numbers they are shown were from the first worker session in which 67 per cent of workers found no-redistribution to be fair, while 33 per cent found redistribution to be fair. Defining a binary variable D_T equal to 1 when a spectator received treatment, we estimate the treatment effect by running the following.

$$[P_i = 1] = \beta_0 + \beta_1 D_T + \epsilon_i \tag{6}$$

The treatment effect of the information treatment is shown in Table Treatment. The effect of being in the information treatment, β_1, was estimated to be -0.007(0.84). To allow for non-linear effects between the prevalence estimate and treatment effect, we also estimated a difference-in-difference interaction of the effect of being in the

[6] The psychology literature on the self in social cognition (Alicke, Dunning and Krueger, 2005) offers a compelling defence for the assumption that prevalence estimates and paternalist motivation are not negatively correlated. Dissonance minimization will lead norm-seeking spectators with a strong dedication to their fairness view to avoid information that lowers the estimate of how many share their view, biasing their estimates of how many share their view upwards. This is a common general explanation for our tendency to overestimate the commonness of our own traits, and is referred to as projection bias. This mechanism will also make the correlation between estimates of how many share a view and willingness to implement a fairness view, positive.



The bottom dark (blue) bars represent the portion of spectators who are willing to pay to implement their redistribution preference, while the top light (magenta) bars are spectators choosing not to implement their preferred redistribution. The total bars show spectators with a prevalence estimate within the bin on the horizontal line. The vertical lines depict the true prevalence of redistribution preferences among the workers.

Figure 2: Histogram of spectators' prevalence estimates of workers fairness view and spectator implementation rate.

Table 3: Estimation of (Reg2): Treatment effect of being informed about the true prevalence of fairness view.

	Dependent variable:
	$[P_i = 1]$
D_T	-0.007
	(0.036)
Constant	0.684^{***}
	(0.027)
Observations	672
R^2	0.0001
Adjusted R^2	-0.001
Residual std. error	0.467 (df = 670)
F statistic	0.041 (df = 1; 670)
Note:	*p<0.1; **p<0.05; ***p<0.01

treatment group and having a prevalence estimate higher than the true prevalence among the workers, $D_{\alpha_i(s)>\alpha_W(s)}$, as follows.

$$[P_i = 1] = \beta_0 + \beta_1 D_{\alpha^i(s)>\alpha_W} + \beta_2 D_T + \beta_3 [D_T \times D_{\alpha_i(s)>\alpha_W}] + \epsilon_i \tag{7}$$

The results from estimating (diffInDiff) are shown in Table Tab: Projection Bias Effect, Treatment Effect, and Their Interaction Effect among Treatment Participants that Overly Projected Prevalence. The parameters $\beta_1, \beta_2, \beta_3$ where estimated to be $[\beta_1 = 0.046(0.43), \beta_2 = 0.064(0.24), \beta_3 = -0.100(0.21)]$. The difference in the treatment effect for participants with high or low prior prevalence estimates, β_3, is almost estimated to be zero; this implies there was no difference in the effect of being informed about the true prevalence on the probability of implementing for spectators with high or low prior prevalence estimates. In other words, the effect of being positively or negatively surprised about the prevalence of one's fairness view on the decision to implement equals zero. This suggests that being informed has no effect on participants' decision to implement. If the true prevalence was relevant to participants with non-paternalist motivation being informed that their prior belief of prevalence was deflated or inflated, it should have opposite effects, giving a negative β_3 estimate. The absence of differences between these groups is strong evidence that information about the true prevalence was irrelevant to the implementation decision. The treatment effect is robustly zero considering subsamples of spectators with particularly high prior estimates of the commonness of their own fairness view; in other words, there was no treatment effect for individuals with prior beliefs of prevalence particularly far from the actual prevalence. We denote the actual prevalence of workers' fairness views as $\alpha_W(s)$. We found highly insignificant treatment effects with varying direction of treatment effect when running (Reg2) for the subsample of spectators with prevalence estimate biases $[\alpha_i(s) - \alpha_W(s)]$ larger than 20, 30 and 40.

Table 4: Estimation of (diffInDiff): Difference-in-difference regression of high prior prevalence estimates and information treatment.

	Dependent variable:
	$[P_i = 1]$
$D_{\alpha_i(s)>\alpha_w(s)}$	0.046
	(0.058)
D_T	0.064
	(0.067)
$D_T \times D_{\alpha_i(s)>\alpha_w(s)}$	-0.100
	(0.080)
Constant	0.652^{***}
	(0.049)
Observations	672
R^2	0.002
Adjusted R^2	-0.002
Residual std. error	0.467 (df = 668)
F statistic	0.546 (df = 3; 668)
Note:	*p<0.1; **p<0.05; ***p<0.01

These results can be seen in Tables App1 – App3 in Appendix sec: Effects Associated with Sizes of Projection Bias.

Furthermore, we tested whether the distribution of spectators implementing conditional on the prevalence estimates is equal for the control and treatment distributions. We conducted the Kolmogorov–Smirnov test on the distributions of spectators' willingness to implement conditional on their prevalence estimate for the control and treatment samples. We found the distributions to be equal for both the control and treatment groups (p-value of 0.689), which is a strong indicator that being informed about the true prevalence had no effect on willingness to implement their preferred distribution; these results are reported in Appendix sec: Control Group versus Treatment Group.

Summary of main findings

Spectators systematically overestimated the prevalence of their own fairness view. Assuming spectators' utility of implementing the state they found fair is a continuous distribution, implies that a portion of spectators had only marginal positive utility when choosing to implement their view or keeping their two cents. About two-thirds of spectators choose to implement the redistribution they found fair. This is true for participants finding they share the minority view of the workers, viewing redistribution as fair, or the majority view, finding no-redistribution fair. This would imply any changes in spectators' utility should change some spectators' utility of implementing the redistribution they find fair from positive to negative. Hence, if the information that the actual prevalence of people sharing the decision maker's fairness view was relevant, and we assume utility from implementing to be a continuous distribution, it should tip some individuals to not implement. However, there was no difference in the numbers of spectators implementing between the control group and

the information treatment, nor any correlation or distributional difference between prevalence estimates and willingness to implement or differences in distributions. We conclude from this that within the established definitions in the theory section, we confirm the presence of paternalist motivation, and do not find evidence to indicate any non-paternalist motivation among the spectators in our sample.

5. Conclusion

We have found spectators' willingness to implement the redistribution alternative they found fair, to be uncorrelated with their estimate of how many people share their fairness view. Furthermore, spectators have an upward bias in their estimates of how many among the affected workers shared their view of what the fair redistribution option is, but being informed about the true prevalence had no effect on their willingness to implement their fairness view.

An implication of our finding is that efforts guided by a paternalist motivation to impose states upon others may have socially negative consequences from a utilitarian perspective; such efforts might lower the utility of the persons affected if people have a lower utility of living under institutions adhering to fairness views they do not hold themselves. An applied example of this problem is the question of whether to give school teachers greater autonomy over choosing the curriculum; our result is supportive that they will put in an effort to do so, but will not take into consideration political sensitivity beyond what issues the teachers themselves find to be fair. However, education might very well be better if the teacher is given autonomy and thus receives a utility bonus from teaching a curriculum she believes to have the fair message.

Our finding poses new questions for future research. Under paternalist motivation there exists a trade-off between voluntary participation to change outcomes for others

to states they do not find fair and the alternative of no-contribution. Fully analyzing the cost and benefit of delegation under paternalist motivation remains both an empirical and theoretical subject open for new research. Empirical questions of interest are whether the finding generalizes beyond fairness norms to political issues, religious preferences or lifestyle choices, and how distributions and strength of preferences of the affected are relevant for implementation decisions. Furthermore, the design and theory apply a binary alternative to implementation. How another default outcome, when the decision maker omits to implement, is a very much applied problem for inquiry; this enlightens the question of the trade-off between omission versus the risk of erroneous commission. Theoretical insights into the desirability of implementation under different degrees of paternalist motivation and different preference structures of the affected could also be developed.

Acknowledgments

We are grateful to the FAIR Centre and the Department of Economics for financial and organizational support in conducting the experiment. The authors gratefully acknowledge support from the Research Council of Norway FRIPRO project "Understanding Paternalism" grant number 262636 and the FAIR Centre through the Research Council of Norway grant 262675. The authors would like to thank Bertil Tungodden for excellent comments and encouragement. We would also like to thank Mathias Philip Ekström, Eirik Andre Strømland, Ole-Andreas Elvik Næss, Jonas Tungodden, seminar participants at FAIR and the Norwegian Meeting of Economists 2017 for comments. We would also like to thank Sebastian Fest for research assistance.

Appendix A. Appendix A: Further results

Appendix A.1. Effects associated with sizes of projection bias

Appendix A.1.1. With projection biases larger than 20

Table A.5: Correlation between projected prevalence and spectators' willingness to implement their fairness views: Subsamples with projection biases larger than 20

	Dependent variable:
	$[P_i = 1]$
$\alpha_i(s)$	0.002
	(0.003)
Constant	0.468**
	(0.211)
Observations	180
R^2	0.004
Adjusted R^2	−0.001
Residual std. error	0.479 (df = 178)
F statistic	0.765 (df = 1; 178)
Note:	*p<0.1; **p<0.05; ***p<0.01

Appendix A.1.2. With projection biases larger than 30

Table A.6: Correlation between projected prevalence and spectators' willingness to implement their fairness views: Subsamples with projection biases larger than 30

	Dependent variable:
	$[P_i = 1]$
$\alpha_i(s)$	0.006
	(0.004)
Constant	0.178
	(0.325)
Observations	85
R^2	0.025
Adjusted R^2	0.013
Residual std. error	0.477 (df = 83)
F statistic	2.139 (df = 1; 83)
Note:	*p<0.1; **p<0.05; ***p<0.01

Appendix A.1.3. With projection biases larger than 40

Table A.7: Correlation between projected prevalence and spectators' willingness to implement their fairness views: Subsamples with projection biases larger than 40

	Dependent variable:
	$[P_i = 1]$
$\alpha_i(s)$	0.010
	(0.012)
Constant	−0.203
	(0.999)
Observations	34
R^2	0.021
Adjusted R^2	−0.010
Residual std. error	0.496 (df = 32)
F statistic	0.680 (df = 1; 32)
Note:	*p<0.1; **p<0.05; ***p<0.01

Appendix A.2. Kolmogorov–Smirnov tests

The Kolmogorov–Smirnov test is based on a measure of distances between two cumulative distribution functions (CDF). We investigate whether the empirical CDF of participants' prevalence estimates have sufficient distance such that they are likely to be generated from different distributions. These are the distributions graphically presented in Fig. fig: Histogram of Aplha. The p-value of the Kolmogorov–Smirnov test estimates the probability that two observed samples are drawn from the same distribution.

Appendix A.2.1. All spectators

We test the equality of the distributions between those who were willing to pay and those who were not, using Kolmogorov–Smirnov tests. The p-value is 0.9672.

Appendix A.2.2. Control group versus treatment group

We test the equality of the distributions between those who implemented in the control group and those who implemented in the treatment group, using Kolmogorov–Smirnov tests. The p-value is 0.689.

Appendix A.2.3. Control group

Within the control group, we test the equality of the distributions between those who implemented and those who did not, using Kolmogorov–Smirnov tests. The p-value is 0.9201.

Appendix A.2.4. Control group: Prefer no-distribution

Within the subsample in the control group preferring no-distribution, we test the equality of the distributions between those who implemented and those who did not, using Kolmogorov–Smirnov tests. The p-value is 0.8231.

Appendix A.2.5. Control group: Prefer distribution

Within the subsample in the control group preferring distribution, we test the equality of the distributions between those who implemented and those who did not, using Kolmogorov–Smirnov tests. The p-value is 0.8943.

Appendix A.2.6. Treatment group

Within the treatment group, we test the equality of the distributions between those who were willing to pay and those who were not, using Kolmogorov–Smirnov tests. The p-value is 0.9862.

Appendix A.2.7. Treatment group: Prefer no-distribution

Within the subsample in the treatment group preferring no-distribution, we test the equality of the distributions between those who were willing to pay and those who were not, using Kolmogorov–Smirnov tests. The p-value is 0.6823.

Appendix A.2.8. Treatment group: Prefer distribution

Within the subsample in the treatment group preferring distribution, we test the equality of the distributions between those who were willing to pay and those who were not, using Kolmogorov–Smirnov tests. The p-value is 0.2267.

Appendix B. Appendix B: Experiment overview

Each spectator makes a decision for one round of work by one pair of workers. Each worker works four separate rounds and is paired with a different worker and different spectator in each round. Participants are randomly allocated as workers or to one of the two spectator treatments.

Appendix B.1. Sequence for workers

- Stage 1: The workers are asked whether they found redistribution or no-redistribution fair in a hypothetical scenario.
- Stage 2: Workers are then allocated to pairs of two. Workers perform a work task consisting of identifying what number is next to a letter on a list of letter–number combinations. Each worker draws a random price and earned the number of correctly identified letters multiplied by his or her drawn price.
- Stage 3: Workers are paid according to the choice of one spectator.

For each pair of workers, stages 2 and 3 are repeated for four rounds. In each round, the workers are paired with a new partner and the pair is assigned to a new unique spectator. Each spectator decides whether to implement her preferred redistribution preference upon one pair of workers for one round.

Appendix B.2. Sequence for control-group spectators

- Stage 1: Spectators are elicited for their third-party redistribution preferences.
- Stage 2: Spectators are elicited for their beliefs regarding how many of the workers shared their third-party preferences, $\alpha^i(s)$.
- Stage 3: Spectators choose whether to pay USD 0.02 to implement their preferred redistribution option.

Appendix B.3. Sequence for treatment-group spectators

- Stage 1: Spectators are elicited for their third-party redistribution preferences.
- Stage 2: Spectators are elicited for their beliefs regarding how many of the workers share their third-party preferences, $\alpha^i(s)$.

- Stage 3: Spectators are informed about the empirical prevalence of their redistribution preference among the workers, α_W.
- Stage 4: Spectators choose whether to pay USD 0.02 to implement their preferred redistribution option.

Appendix C. Appendix C: Experimental instructions

All text in italics is left out of the experiment and only included for a reader overview. Screens indicate when the program will change text.

Appendix C.1. Worker instructions

Appendix C.1.1. Screen 0: Instruction on M:Turk

Task Link Instructions (Click to expand)

Thank you for your participation in this task. Please read all instructions carefully. The results from this task will be used in a research project at the Norwegian School of Economics. Participation in the study is completely voluntary.

You are free to decline to participate, or to end participation at any time and for any reason.

Your will remain anonymous throughout the task. None of the information collected can be traced back to individual participants. We will only use your participant ID to assign payments and to check that you have not participated in this task before. The duration of the task is approximately 5 min.

If you have any questions regarding this task, please contact xianwen.chen@gmail.com.

To verify that you have actually completed the task, you are required to enter a unique participant ID below. You will receive your participant ID at the end of the task, following the link below.

Task link: *Link to the on-line task*

Provide the participant ID here: ____

Appendix C.1.2. Screen 1: Introduction

Thank you for your participation in this task. Please read all instructions carefully. The results from this task will be used in a research project at the Norwegian School of Economics. Participation in the study is completely voluntary.

You are free to decline to participate, or to end participation at any time and for any reason.

Your will remain anonymous throughout the task. None of the information collected can be traced back to individual participants. We will only use your participant ID to assign payments and to check that you have not participated in this task before.

The duration of the task is approximately 5 min.

If you have any questions regarding this task, please contact xianwen.chen@gmail.com.

Click the ">>" button to indicate that you have read and understand the above information and that you agree to participate in this study.

Appendix C.1.3. Screen 2: Hypothetical Scenario

Assume two workers have been completing an identical task. The task is identifying the number on a list that is next to a given letter. The worker gets paid per correctly identified word. Each worker gets paid a separate randomly drawn price. The price can be any whole number from 1 to 10 cents.

Elicit hypothetical preferences. Which of the payment options do you find to be the most fair option?

1. No-redistribution: Each worker is paid separately for their work. In other words the workers get paid for the number of words they identified times the

price they are randomly assigned.

2. Redistribution: The total earnings of the two of workers are divided equally among the workers. In other words each worker gets paid the sum of the payments of the two workers divided by two.

Appendix C.1.4. Screen 3: Work task

You are now going to perform a letter-number decoding task equal to that described in the previously described scenario. A sequence of letters with corresponding numbers will be displayed on the screen. You should write the number corresponding to the given letter in the box below the sequence.

An example of the task is provided below *(Figure fig: example letter identifier list).* You should type the number that corresponds to the letter O, which in this case is 47.

A new sequence will be displayed directly below the first. You will not know whether your answers are correct until the end of the task. The task will last for 60 seconds. Remaining time will be displayed at the top of the page.

After completing the task a price between 1 and 10 cents per word will be drawn at random. Another experimental participant will choose whether your earnings will be redistributed with another worker, or whether you will receive payment equal to your earnings.

Appendix C.1.5. Screen 4

You will now perform the letter-number decoding task.

When you are ready, press ">>" to start the task.

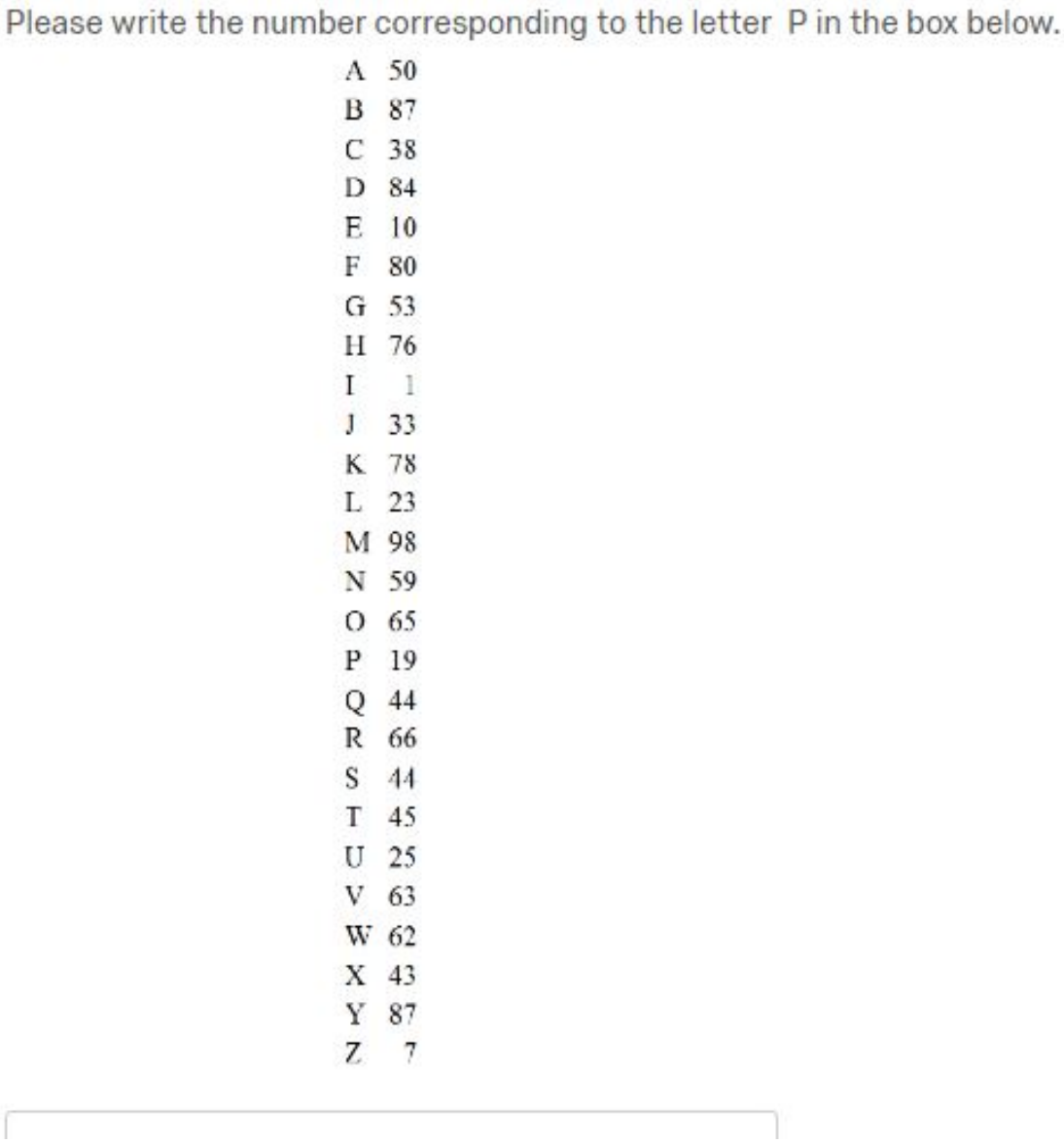


Figure C.3: Example of letter identifier list.

Appendix C.1.6. Screen 5

Twenty letter identifier list will appear for the workers. Each worker is given one minute to complete as many tasks as possible.

Appendix C.1.7. Screen 6

You solved *(number of correctly identified numbers).*

You will now perform the letter-number decoding task one more time.

When you are ready, press ">>" to start the task.

Appendix C.1.8. Screen 7

Twenty letter identifier list will appear for the workers. Each worker is given one minute to complete as many tasks as possible.

Appendix C.1.9. Screen 8: Information of payment

You solved *(number of correctly identified numbers in the second round)* this round at the drawn price of *(randomly drawn price in the second round)* cents.

In addition, you solved *(number of correctly identified numbers in the first round)* in the first round at the drawn price of *(randomly drawn price in the first round)* cents.

Thank you for participating in the study. Please remember that your participant ID is *(randomly generated participant ID).*

You will receive your payment within 2 weeks.

When you are ready, it is very important that you press ">>" to end the survey!

Please remember to submit your participant ID *(randomly generated participant ID)* in Amazon MTurk!l

Appendix C.2. Spectator session: control-group

Appendix C.2.1. Screen 0: Instruction on M:Turk

Survey Link Instructions (Click to expand)

Thank you for your participation in this survey. Please read all instructions carefully. The results from this survey will be used in a research project at the Norwegian School of Economics. Participation in the study is completely voluntary.

You are free to decline to participate, or to end participation at any time and for any reason.

Your will remain anonymous throughout the survey. None of the information collected can be traced back to individual participants. We will only use your participant ID to assign payments and to check that you have not participated in this survey before.

The duration of the survey is approximately 5 min.

If you have any questions regarding this survey, please contact xianwen.chen@gmail.com.

To verify that you have actually completed the survey, you are required to enter a unique participant ID below. You will receive your participant ID at the end of the survey, following the link below.

Survey link: *Link to the on-line survey*

Provide the participant ID here: ____

Appendix C.2.2. Screen 1: Introduction

Thank you for your participation in this survey. Please read all instructions carefully. The results from this survey will be used in a research project at the Norwegian School of Economics. Participation in the study is completely voluntary.

You are free to decline to participate, or to end participation at any time and for any reason.

Your will remain anonymous throughout the task. None of the information collected can be traced back to individual participants. We will only use your participant ID to assign payments and to check that you have not participated in this task before.

The duration of the survey is approximately 5 min.

If you have any questions regarding this survey, please contact xianwen.chen@gmail.com.

Click the ">>" button to indicate that you have read and understand the above

information and that you agree to participate in this study.

Appendix C.2.3. Screen 2: Hypothetical Scenario

Same as Screen 2 of workers.

Appendix C.2.4. Screen 3: Real Scenario

There has been performed an experiment like the one described on the M:Turk platform with participants like yourself. Worker participants completed the task as described and drew a random price between 1 and 10 cents. Before completing the work task the workers were asked which of the payment options, redistribution or no-redistribution they found to be the fair option.

Elicit α: for participants with third party preferences for redistribution. Among 100 worker participants how many do you think find the redistribution option fair?

Elicit α: for participants with third party preferences for no-redistribution. Among 100 worker participants how many do you think find the no-redistribution option fair?

Appendix C.2.5. Screen 4A: Elicit real 2 cent preferences for participants with third party preferences for redistribution

You are to decide how payment should be done for one pair of worker participants. You can pay 2 cents of your 1.11 dollar participation earning to implement the distribution you previously stated you found to be the fair option: the redistribution option.

I choose to:

1. Pay 2 cents to implement the redistribution option.
2. Not pay 2 cents. The no-redistribution option will then be implemented.

Appendix C.2.6. Screen 4B: Elicit real 2 cent preferences for participants with third party preferences for no-redistribution

You are to decide how payment should be done for one pair of worker participants. You can pay 2 cents of your 1.11 dollar participation earning to implement the distribution you previously stated you found to be the fair option: the no-redistribution option.

Would you like to pay 2 cents to implement the no-redistribution option?

I choose to:

1. Pay 2 cents to implement the no-redistribution option.
2. Not pay 2 cents. The redistribution option will then be implemented.

Appendix C.2.7. Screen 5: End

Thank you for participating in the study. Please remember that your participant ID is *(randomly generated participant ID)*.

You will receive your payment within 2 weeks.

When you are ready, it is very important that you press ">>" to end the survey! Please remember to submit your participant ID *(randomly generated participant ID)* in Amazon MTurk

!

Appendix C.3. Spectator session: Treatment-group

For the treatment-group spectator session, everything is identical as in the control-group spectator session, except that an additional screen is added between Screen 3 and Screen 4. In the new screen, the following treatment text is displayed. Treatment Screen: Revelation of empirical α

Treatment Screen: Revelation of empirical α for participants with third party preferences for redistribution:

Among 100 worker participants, XX found the redistribution option to be the most fair option.

Treatment Screen: Revelation of empirical α for participants with third party preferences for no-redistribution:

Among 100 worker participants, XX found the no-redistribution option to be the most fair option.

Appendix D. References

Alicke, Mark D, David A Dunning and Joachim Krueger. 2005. *The self in social judgment*. Psychology Press.

Buser, Thomas, Louis Putterman and Joel J van der Weele. 2016. "Gender and Redistribution: Experimental Evidence." *Discussion paper, Tinbergen Institute, UvA.* .

Cappelen, Alexander W, Astri Drange Hole, Erik Ø Sørensen and Bertil Tungodden. 2007. "The pluralism of fairness ideals: An experimental approach." *American Economic Review* 97(3):818–827.

Chen, Xianwen and Øivind Schøyen. 2017. "Costly implementation of third party preferences with non-paternalistic motivation." *American Economic Association RCT Registry* (2310).

Coons, Christian and Michael Weber. 2013. *Paternalism: Theory and practice*. Cambridge University Press.

Durante, Ruben, Louis Putterman and Joël Van der Weele. 2014. "Preferences for redistribution and perception of fairness: An experimental study." *Journal of the European Economic Association* 12(4):1059–1086.

Rosati, Connie S. 2006. "Moral motivation." *The Stanford Encyclopedia of Philosophy* .

Ross, Lee, David Greene and Pamela House. 1977. "The "false consensus effect": An egocentric bias in social perception and attribution processes." *Journal of Experimental Social Psychology* 13(3):279–301.

Rothschild, David and Neil Malhotra. 2014. "Are public opinion polls self-fulfilling prophecies?" *Research & Politics* 1(2).

Thaler, Richard H and Cass R Sunstein. 2003. "Libertarian paternalism." *The American Economic Review* 93(2):175–179.