Social Roles

Avner Seror

amse
école d'économie d'aix-marseille
aix-marseille school of economics

Working Papers / Documents de travail

Social Roles

Avner Seror

WP 2021 - Nr 34

Social Roles

Avner Seror*
Aix Marseille School of Economics

May 25, 2021

Abstract

In this paper, I introduce a workable dynamic utility model on the interplay between economic actions and social roles. I model both how economic actions are embedded in social roles, and how social roles reciprocally feed back into preferences and affect economic outcomes. I also consider a set of policy interventions aimed at breaking social roles when they deteriorate economic outcomes.

JEL Classification Numbers: D1, D7, D9, H3, J15, J16, J7, Z1
Keywords: Social Roles, Identity, Endogenous Preferences, Gender, Discrimination

1 Introduction

Social roles have been central to the functioning of all economies since the Paleolithic Era. Historically, social roles depend on gender, age, kinship, race, ethnicity or religion and structure economic production and exchange (Sahlins (2017 [1974])). Sociologists have long recognized their importance in economic decisions (Granovetter (1985)). By contrast, rooted in the neoclassical tradition, most economic models operate under the assumption that agents are under-socialized, and it is only recently that social roles have been conceptualized in formal models (Akerlof and Kranton (2000), Montgomery (2004), Shayo (2009), Bordalo et al. (2016)). While these approaches are relevant in many instances,

*avner.seror@univ-amu.fr, Aix-Marseille Univ., CNRS, AMSE, Marseille, France, 5-9 Boulevard Maurice Bourdet, Marseille, 13001, France. This work was supported by the French National Research Agency Grant ANR-17-EURE-0020, and by the Excellence Initiative of Aix-Marseille University - A*MIDEX. All errors are my own.

little has been done to analyze social roles as a global phenomenon that interacts with the development of society.

In this paper, I introduce a workable dynamic utility model on the interplay between economic actions and social roles. I model both how economic actions are embedded in social roles and how social roles reciprocally feed back into internalized preferences and affect economic outcomes. I start the analysis with a static version of the model. There is a finite set of agents playing a contribution game.[1] The agents have different abilities to perform the tasks at hand and must choose a contribution. An agent's utility depends on everyone's contributions, but also on his beliefs about his own social role and the social roles of others. Each agent can impose his beliefs on the social roles of others through a costly punishment technology. I find that there is a unique equilibrium, which reflects a key interaction between social roles and contribution decisions: when it is very important to an agent that others abide by their social role, he relies on the punishment technology to make others' actions conform to his own beliefs regarding their social roles.

To see the reasoning behind the static model, consider the following example. A society is divided between men and women. Some men believe that their social role is to be breadwinners. They also believe that the social role of women is to stay home. Depending on the value to them of holding these beliefs, men might use domestic violence, discrimination or harassment to keep women home, so that they end up not directly contributing to the economy. This static model illustrates how social roles affect standard economic decisions. It can be applied to many cases: gender roles enforced through domestic violence, harassment, and discrimination, sexism in corporate culture and educational choices, and social exclusion of racial, sexual, ethnic or religious minorities. Through two extensions to market games and self-nomination games, the model can also be applied to labor market discrimination, participation in strategic decisions (e.g. self-nomination for leadership positions), and other outstanding instances of social roles affecting economic outcomes.

Although I start with a static model, the main results of this analysis come from the dynamic model where the static game is indefinitely repeated. At the end of each round of the game, the agents can revise their beliefs on social roles. The process of belief formation is modeled as an internal design mechanism where the agents adopt the best possible beliefs on social roles given their experienced utility (Kahneman, Wakker and Sarin (1997)). I find that the agents' optimal beliefs will reflect the optimal contributions of the

[1]The analysis can equally be applied to other standard game settings such as market games and self-nomination games. See Section 6.

previous round of the game. This creates a fundamental interdependence between social roles and economic decisions. As the agents make optimal contribution decisions, they revise their beliefs on social roles, which feed back into their preferences and affect their future contribution decisions. The resulting joint dynamics of social roles and economic contributions display two types of stationary states.

In the first equilibrium, agents' contributions and social roles reflect their ability to perform the tasks at hand. This equilibrium can be reached in two cases that I characterize. First, this equilibrium is reached when, for any agent, the importance of others' abiding by their social roles is low. In that case, no one is punished initially. Social roles then gradually change to reflect the agents' actions, up to a point where both the equilibrium actions and the social roles correspond to the agents' abilities.

Importantly, I find that this long-run equilibrium can also be reached when there is strong initial disagreement between agents on their respective social roles that triggers punishment from an arbitrary number of them. Indeed, although punishment may initially deter several agents from contributing as they would like, it harmonizes the agents' beliefs on their respective social roles. Hence, after an initial period of punishment, the agents' contribution decisions will more closely reflect the beliefs held by others on their social role. Yet their contributions remain influenced by their abilities. Thus, after the second round of the game, social roles and contribution decisions can evolve gradually to reflect the agents' abilities. This dynamic outcome arises when others' abiding by their social role assumes an intermediate level of importance, so that only large deviations from social roles are punished. Small deviations are not punished and arise endogenously after the first round of the game.

In the second equilibrium, some agents are punished initially and continue to be punished if they deviate from their assigned social role in subsequent rounds of the game. As a result, the long-run equilibrium is such that some agents will be perceived by everyone, including themselves, as having social roles that do not match their abilities. This equilibrium is reached when the importance of others' abiding by their social role is great.

I then project this general abstract framework onto two case studies. First, the model is applied to the evolution of gender roles and economic outcomes. The model accounts for two key features frequently addressed in the literature on the evolution of female labor force participation: i) men and women hold beliefs on their respective social roles and ii) men can impose various forms of punishment on women when their behavior deviates from what men consider women's gender role. In this context, I study whether different initial

conditions of the model could generate distinct dynamic paths. I also apply the model to the social roles attributed to Black and White workers in the American South. In that case, the model describes how greater use of slave labor in the American South affected both the evolution of beliefs on the social role of Black workers and development outcomes.

Social roles also have major welfare implications. Depending on beliefs on social roles, I find that the long-run equilibrium where punishment persists is Pareto dominated by the equilibrium where there is no punishment. This is the case, for example, when some individuals are deterred from contributing by a threat of punishment. Every agent loses from a lower aggregate contribution, including those that implement the punishment threat. In such cases, there is room for policy interventions that can redirect the dynamics toward the Pareto optimal long-run equilibrium.

I establish two key policy implications. First, imposed temporary quotas that lead to the victims of punishment being more heavily represented in organizations can put an end to punishment threats. When a large enough number of agents are subject to punishment, punishing them is too costly as it negatively impacts the group's production. Quotas create a *window of opportunity* for oppressed groups to realize their economic potential and break inefficient beliefs on social roles. Second, I find that laws and social movements that impose meaningful constraints on the perpetrators of violence can also create such windows of opportunity. For example, both the #MeToo and the #BlackLivesMatter movements imposed high costs on the perpetrators of various forms of violence aimed at enforcing social roles. The #MeToo movement temporarily increased the cost of sexual harassment and sexual abuse by making women's allegations public. Similarly, the #BlackLivesMatter movement, by denouncing instances of police violence against Blacks, raised the cost of racial violence for police officers. Both movements may have offered a window of opportunity, likely increasing female labor force participation and decreasing the social exclusion of Blacks in the United States.

This paper contributes to several strands of the literature. Principally, it contributes to the large and multifaceted economic literature on social influences on preferences and economic outcomes, in two main ways.[2] Firstly, it is typically assumed that when individ-

[2]One major contribution to the study of social influences on preferences and economic outcomes is Akerlof and Kranton (2000). Identity has been modeled as preferences of individuals when they wish to associate themselves with different groups by Atkin, Colson-Sihra and Shayo (2021); Sambanis and Shayo (2013); Shayo (2009, forthcoming). An alternative approach is to model identities as beliefs, see Bénabou and Tirole (2002, 2003, 2004, 2011*b*). Finally, the approach of Bordalo et al. (2016) considers that distinctive group characteristics are used to build heuristics in probability judgments.

uals associate themselves with groups, their preferences are such that they compare their own behavior with the average behavior in these groups (Shayo (2009)). I assume that how others behave also matters when individuals seek to associate themselves with different groups, and that internalized beliefs on social roles are the outcome of an equilibrium involving many agents. For example, for a married man to identify as a breadwinner, not only must his economic actions conform to his social role (e.g. working) but his wife's actions too must conform to what he believes is her social role (e.g. staying home).[3] I show that this approach is relevant, as a model that does not account for these aspects of individual preferences cannot explain the persistence of inefficient beliefs on social roles.

Secondly, on the evolution of preferences, for which several mechanisms have been advanced,[4] I propose a novel approach that builds on Kahneman, Wakker and Sarin (1997) where preferences on social roles respond optimally to the experienced utility of the agents.[5] I find that this approach is the most consistent with empirical and experimental regularities, showing that when preferences on social roles are only impacted by agents' decision utility, then inefficient beliefs on social roles cannot persist in the long run. Finally, I demonstrate that my approach extends to more complicated cases where the evolution of preferences is influenced by both experienced utility and decision utility.

This paper also contributes to the growing literature on gender economics, in three ways. First, this framework provides a grid for interpreting the long-term persistence of gender roles in households, firms, schools, and society (Alesina, Giuliano and Nunn (2013), Jayachandran (2015)). Second, I show that the joint evolution of internalized gender roles and economic outcomes could explain various differences between the preferences of men and women reported in the literature.[6] Finally, this model explains the widely documented

[3]My approach is rooted in the long tradition in sociology that sees individuals as embedded in social roles that are internalized through preferences (Granovetter (1985), Montgomery (1998), Montgomery (2004), Stryker and Burke (2000)).

[4]The choice of preferences can be driven by parents' investments in intergenerational transmission (Bisin and Verdier (2001), Tabellini (2008), Bisin and Verdier (2011)), anticipatory utility, self-esteem and commitment (Bénabou and Tirole (2011*a*)), self-esteem (Akerlof (2017)), world-views that shape judgment (Bernheim et al. (2021)), cognitive dissonance (Oxoby (2003), Oxoby (2004), Rabin (1994), Akerlof and Dickens (1982)), be the result of an evolutionary process (Alger and Weibull (2013), Robson and Samuelson (2011)) or interact with the evolution of institutions (Bowles (1998), Alesina and Giuliano (2015), Bisin and Verdier (2017), Besley and Persson (2019), Besley (2020), Bisin et al. (2021)).

[5]This approach is also closely related to the theory of institutional change advanced by Bisin and Verdier (2017).

[6]For example, men and women have been shown to have different attitudes toward risk (Croson and Gneezy (2009) and Eckel and Grossman (2008)) or competition (Niederle and Vesterlund (2007)). Bertrand (2011) reviews the related literature.

pervasive effects of internalized gender roles both on individual behaviors (Coffman (2014), Reuben, Sapienza and Zingales (2014), Bursztyn, Fujiwara and Pallais (2017)) and on key development outcomes such as women's labor force participation, political representation or educational choices and performance.[7]

Finally, I establish several directly testable predictions in contribution games, market games, and self-nomination games regarding policy interventions aimed at breaking internalized beliefs on social roles that negatively impact economic behaviors. Hence, this paper also contributes to the experimental literature that seeks both to document the impact of social preferences on economic behaviors and to study how they can be changed in the long run (Coffman (2014), Bohnet, van Geen and Bazerman (2016) , Bursztyn, Fujiwara and Pallais (2017), Bursztyn, González and Yanagizawa-Drott (2020), and Bursztyn, Egorov and Fiorin (2020)).

The rest of the paper is organized as follows. The next section introduces the static model. Section 3 introduces the dynamic setup, while Section 4 discusses the interplay between the evolution of social roles and development outcomes. Section 5 presents a welfare analysis and the policy implications of the model. In Section 6, I extend the model to self-nomination games and market games, while Section 7 concludes. All the proofs are contained in the Appendix.

2 The Static Model

I first present a static model that introduces the main economic forces.

2.1 Agents and Monetary Payoffs

There is a finite population of agents, $\mathcal{N} = \{1, \ldots, n\}$. In the first stage, the agents play a game and choose a behavior. Although applied here to a public good game, the model and the results extend to a broader set of games, including self-nomination games and market games.[8] The game has two stages. In the first stage, agent i chooses a contribution effort $a_i \geq 0$. Agent i has a known ability $\theta_i \in [0, 1]$, so her contribution to the common pool

[7] See, among others, Bertrand and Mullainathan (2004) on labor market discrimination, Bertrand, Kamenica and Pan (2015) and Olivetti and Petrongolo (2016) on the gender wage gap, Gilardi (2015) on women's political representation and Niederle and Vesterlund (2010) on women's educational performance. Inglehart and Norris (2003) and Bertrand (2020) provide general overviews.

[8] See Section 6.

is $\theta_i a_i$. Exerting an effort a_i requires a quadratic cost $a_i^2/2$. In the second stage, agent i chooses to impose a punishment $p_{ij} \geq 0$ on agent $j \neq i$. Hence, the monetary payoff of agent i can be written as

$$x(a_i, \mathbf{a_{-i}}) = \sum_{j \in \mathcal{N}} \theta_j a_j - \frac{a_i^2}{2} - \sum_{j \in \mathcal{N}} \sum_{i \in \mathcal{N}} p_{ij}. \tag{1}$$

2.2 Social Roles and Identities

I now define the building blocks of the proposed model.

Social Identities. Suppose that there exist only two social identities $s \in \{1, 2\}$. The social identity of agent i is denoted $s_i \in \{1, 2\}$. I assume that there are n_1 agents of type 1 and n_2 agents of type 2. For example, agents self-identify as either men or women. Given this paper's objectives, I abstract from decisions relative to social identification and identity formation.[9]

Social Roles. Following Granovetter (1985), I assume that agents have social roles in economic actions. For example, a man may believe that his male social role is to contribute to the common pool while women's role is not to contribute. Each agent is characterized by a vector of beliefs on social roles that she attributes to herself and to others $\mathbf{r_i} = \{a_j(i)\}_{j \in \mathcal{N}}$, where $a_j(i) \geq 0$ is the contribution effort that should be exerted by agent $j \in \mathcal{N}$, as perceived by agent i. I finally denote $\mathbf{r} = [a_j(i)]$ the square matrix of social roles.

The utility of agent i depends both on his monetary payoff (1) and on the extent to which actions match social roles. I propose the following utility function:

$$u_i(a_i, \mathbf{a_{-i}}, \mathbf{r_i}) = x(a_i, \mathbf{a_{-i}}) - \sum_{j \in \mathcal{N}} \frac{\alpha_{ij}}{2} \mid a_i - a_i(j) \mid^2 - \sum_{j \in \mathcal{N}} \frac{\gamma_{ij}}{2} \mid a_j - a_j(i) \mid^2, \tag{2}$$

with α_{ij} and $\gamma_{ij} \in [0, 1]$, $\sum_{j \in \mathcal{N}} \alpha_{ij} = 1$ and $\sum_{j \in \mathcal{N}} \gamma_{ij} = 1$. The parameter α_{ij} corresponds to the importance to agent i of fulfilling his social role as perceived by agent j. For example, an agent may value adopting an action that is consistent with his own beliefs about his social role. He may also feel compelled to choose an action that fits others' expectations of his social role. By contrast, parameter γ_{ij} corresponds to the importance to agent i of agent j fulfilling her social role $a_j(i)$.

[9] On social identification decisions, see, for instance, Atkin, Colson-Sihra and Shayo (2021) in the case of religious identity, and Seror and Ticku (2020) for sexual identity.

Through the second term on the right-hand side of (2), the utility function accounts for the influence of both *self-image* and *social-image* on economic actions.[10] Through the third term on the right-hand side of (2), the utility function also accounts for the influence on an agent's utility of others' conforming to agent i's beliefs about their social roles. Introducing this social role factor into individual preferences is the key novel feature of my approach relative to the existing economic literature. It applies to a broad array of situations.[11] For example, some men may suffer a loss when women act in a way far removed from what these men consider appropriate female behavior (e.g., not working, choosing "feminine" educational paths, wearing certain clothes, or being thin). More broadly, accounting for this social role factor in preferences could shed light on the effect of internalized kinship structures or social hierarchies on individual behaviors. It also provides a tractable model to study racism, xenophobia, homophobia or transphobia.

2.3 Equilibrium

In this section, I characterize the equilibria of the static model presented in the previous section.

The game consists of two stages. In the first, each agent chooses a contribution effort a_i. In the second stage, each agent decides whether to punish other agents. I denote $a_i^*(\mathbf{r})$ the equilibrium contribution of agent $i \in \mathcal{N}$ and $p_{ij}^*(\mathbf{r})$ the equilibrium punishment administered by agent i to agent $j \in \mathcal{N}$.

Before characterizing the equilibria, I introduce several simplifying assumptions. First, I assume that agents sharing a social identity have the same perception of social roles, so $a_j(i) = a_{s_j}(s_i)$ where $s_k \in \{1, 2\}$ denotes the social identity of agent $k \in \mathcal{N}$. Similarly, I assume that agents sharing a social identity have the same perception of the importance of social roles, so $\alpha_{ij} = \alpha_{s_i s_j}$ and $\gamma_{ij} = \gamma_{s_i s_j}$. Third, I assume that only agents with social identity $s = 1$ can punish other agents. I will denote $\gamma_{12} \equiv \gamma$ in the rest of the paper. This assumption is more demanding, although it enables me to focus on the main moving parts without altering the reasoning of the model. Fourth, I assume that $1 > \alpha_{s_i s_i} n_{s_i} + \alpha_{s_i s_j} n_{s_j}$ for any $i, j \in \mathcal{N}$, $i \neq j$. This assumption simplifies the problem, as it ensures that the

[10] For instance, Abeler, Nosenzo and Raymond (2019) formalize and test a wide range of potential explanations for lying behaviors. The authors demonstrate that honesty can be explained by a combination of self-image and social-image motives.

[11] This specification can be seen as a generalization to a broader set of social roles of the approach of Fehr and Schmidt (1999) to fairness concerns impacting individual preferences.

equilibrium is unique. Finally, I assume that the punishment needs to be paid only once by a type 1 agent, and punishment is administered in bilateral interactions. Hence, I abstract from issues of strategic free-riding in punishment that would naturally arise if punishment occurred in more complex network structures (Bramoullé and Kranton (2007)).

Theorem 1 *Under the previous assumption, there exists a unique equilibrium and a threshold $\tilde{\gamma}$ such that:*

- *If $\gamma < \tilde{\gamma}$, $a_i^*(\mathbf{r}) = \tilde{a}_i(\mathbf{r})$ for any $i \in \mathcal{N}$, with $\tilde{a}_i(\mathbf{r})$ the contribution effort that maximizes (2) and $p_{12}^*(\mathbf{r}) = 0$.*
- *If $\gamma \geq \tilde{\gamma}$, $a_i^*(\mathbf{r}) = \tilde{a}_i(\mathbf{r})$ for any i such that $s_i = 1$ and $a_j^*(\mathbf{r}) = a_2(1)$ for any j such that $s_j = 2$. If an agent of type 1 deviates from action $a_2(1)$, she incurs a punishment $p_{12}^*(\mathbf{r}) > 0$.*

Although a type 1 agent benefits from the contributions of type 2 agents, he also values these agents' conforming to the social role that he assigns them. Hence, when he perceives the importance γ of type 2 agents fulfilling their social role as great, he will punish them if they deviate from it. As a result, when $\gamma > \tilde{\gamma}$, type 2 agents prefer to contribute $a_2(1)$, possibly a lower level than what they would have chosen without the threat of punishment by type 1 agents.

2.4 Motivating Examples

It is useful to have some running examples to fix ideas.

Gender roles. One simple application illustrating the working of the model is gender roles. Suppose that each agent either identifies as a man or as a woman. Men perceive themselves as breadwinners and also believe that the social role of women is to stay home. By contrast, women do not necessarily perceive either themselves or men as exclusive breadwinners. In these settings, men may choose to punish women to force them not to contribute. The punishment can take various forms, from sexual harassment in the workplace to discrimination and domestic violence. The threat of punishment results in an equilibrium where only men are not constrained in their actions and women have to conform to what men expect of them.

Social exclusion of minorities. If some individuals in an organization are homophobic, racist or xenophobic, they may believe that immigrants and people whose religious

beliefs, race or sexual orientation differ from theirs should be socially excluded. For example, acts of xenophobia and racism are related to the beliefs that when jobs are scarce, priority should be given to natives rather than immigrants or that people from different religions or races should have access to a limited set of economic occupations. Such beliefs often create social exclusion and are enforced through discrimination and violence.

Social Hierarchy in the lands of Islam. Historically, Islam decreed a specific social division between Muslims and non-Muslims. From the Quran decree (9.29), "Fight those of the People of the Book who do not [truly] believe in God and the Last Day, who do not forbid what God and his Messenger have forbidden, who do not obey the rule of justice, until they obey the law and agree to submit." In Islamic jurisprudence, this decree was embodied in the Pact of Umar I (634-644), which founded rights and "protection" for non-Muslims (or *Dhimmis*) living under Islamic rule. *Dhimmis* were not allowed to possess weapons or beat Muslims, otherwise, they would not be protected under law. *Dhimmis* were also required to dress differently from Muslims. These rules, when internalized, often led to persecution, extortion, and violence against Jewish and Christian minorities (Bensoussan (2012), Kuran and Lustig (2012)).[12]

These examples clarify the meaning of "social roles". Social roles correspond to beliefs about economic behaviors that are internalized in individual preferences. This model thus formalizes the argument of Granovetter (1985) that action is embedded in social relationships. It also provides a simple way to examine how utility costs arising from internalized social roles can trigger discrimination when some agents' behaviors are inconsistent with what is expected of them by others.

3 Dynamics

The static model shows how the importance agents attach to social roles can influence economic behavior. However, the static setup does not permit analysis of how social roles jointly evolve and influence the overall contribution made by the agents. In this section,

[12]For example, Kuran and Lustig (2012) show that judicial biases against non-Muslim merchants were institutionalized in Ottoman Courts. In another example, de Foucauld (1998), disguised as a Jew, traveled the Moroccan coast in 1883 (after the abolition of the dhimmi status) and noted "They [the Jews] cannot go out without being hit with stones." The author argues that one way Jews were able to mitigate the arbitrary violence they suffered and to engage more safely in commercial activities was to seek the protection of a powerful Muslim or a tribe, a practice called *dehiba*. Relatedly, Johnson and Koyama (2019) give a thorough overview of persecution against Jewish minorities in Europe.

I consider a dynamic generalization of the static model and demonstrate under which conditions social roles lead to inefficiently low participation rates.

3.1 Dynamic model

The dynamic model is a straightforward generalization of the static setup. Agents live for T periods, where T can be either finite or infinite. In each period $t \geq 1$, the agents meet and play a public good game, as outlined in the previous section. The effort of agent i in period t is denoted $a_{i,t} \geq 0$ and the punishment exacted by agent i on agent j is denoted $p_{ij,t}$. Agent i's ability $\theta_i \in [0,1]$ is assumed fixed. I denote $\mathbf{r_{i,t}} = \{a_{i,t}(j)\}_{j \in \mathcal{N}}$ the vector of beliefs on the social roles of agent i in period t and $\mathbf{r_t} = [a_{i,t}(j)]$ the matrix of beliefs on social roles. I denote $a^*_{i,t}(\mathbf{r_{i,t}})$ the equilibrium contribution of agent $i \in \mathcal{N}$ and $p^*_{ij,t}(\mathbf{r})$ the equilibrium punishment administered by agent i on agent $j \in \mathcal{N}$ in period t. Finally, I denote $\mathbf{r}_\infty = [a_{i,\infty}(j)]$ the long-run matrix of social roles.

I assume that in each period, social roles are internalized and acted out. Hence, at the end of any period t, the agents revise their internalized beliefs on these social roles according to their experienced utility in that period (Kahneman, Wakker and Sarin (1997)). The agents choose their beliefs on social roles aiming to maximize their utility:

$$\mathbf{r_{i,t+1}} = \arg\max_{\mathbf{r_i}} u_i(a^*_{i,t}(\mathbf{r}), \mathbf{a^*_{-i,t}}(\mathbf{r})), \tag{3}$$

with $\mathbf{r} = [\mathbf{r_1}, \ldots, \mathbf{r_n}]$.

Agents choose to adopt the beliefs that would have been the most profitable to them in the game just experienced. The internal design of beliefs on social roles is thus retrospective, and reflects an internal design mechanism where each agent evaluates how much better off she would have been had she perceived everyone's social roles differently, including her own.

I focus the exposition of the model on a case where the internal design of beliefs on social roles is only affected by individuals' experienced utility. However, the internal design of beliefs on social roles might also be prospective, as agents could try to adopt beliefs on social roles that maximize their decision utility in future rounds of the game. In Appendix A.2, I build an extension of the model that accounts for both retrospective and prospective concerns in the internal design of beliefs on social roles. I show that provided the internal design of beliefs in period t is in part retrospective, even at an arbitrarily low level, then

all the results of this paper hold. By contrast, when the internal design of beliefs is only prospective, then the dynamics reach a steady state immediately after period 1, where there is no punishment and both beliefs on social roles and equilibrium actions are equal to economic abilities. Hence, a purely prospective model of the formation of beliefs on social roles cannot explain the key empirical and experimental regularities that motivate this theory.

Theorem 2 *In any period t and for any $i, j \in \mathcal{N}$, $a_{j,t+1}(i) = a^*_{j,t}$.*

From the maximization program (3), at the end of each period t, each agent's perception of social roles will reflect the equilibrium actions of all the agents. Indeed, at the end of period t, an agent internalizes the fact that had he perceived the social roles as equal to the equilibrium actions, he would not have suffered a utility cost due to self-image concerns (when his action deviates from what he believes is his own social role), nor a utility cost due to others' deviation from the social role he assigned them.

3.2 Case without punishment

To get a better understanding of the dynamics of social roles, consider first the baseline case where agents cannot punish each other. In this simple case, I establish the following result:

Theorem 3 *There exists a unique equilibrium where in the long run, social roles and actions reflect agent abilities, $a_{i,\infty}(j) = a^*_{i,\infty} = \theta_j$ for any $i, j \in \mathcal{N}$.*

Theorem 3 implies that absent punishment, beliefs on social roles necessarily converge in the long run, reflecting the distribution of economic abilities. Intuitively, from the maximization problem (3), agents of both types adapt their perception of social roles to the equilibrium actions chosen by all the agents. Given that there is no punishment, actions converge to abilities. Hence, so do social roles.

This result is important because it shows that absent punishment, conformism or social image concerns are not sufficient to explain the persistence of social perceptions that constrain economic contributions. As I show next, it is rather the combination of utility cost suffered when others do not conform to their assigned social roles and unconstrained punishment that creates long-run economic inefficiencies.

3.3 Case with punishment

I now introduce one of the main results. I show that when agents suffer identity cost from others' not conforming to their assigned social roles, and when they can use credible punishment threats, inefficient social roles can persist in the long run and constrain economic contributions.

Although the general insight concerning the joint evolution of social roles and economic contributions holds for all parameter values, the analysis in the general case is complex. I will therefore focus on a subset of cases that correspond to the above social role examples and where there is initially a stark conflict between the two types' beliefs on their social roles. I assume that type 1 agents initially believe that the social role of type 2 agents is not to contribute, i.e. $a_{2,1}(1) = 0$. To further simplify, I assume that initially, type 2 agents believe that their contribution should be equal to their ability, $a_{2,1}(2) = \theta_2$.

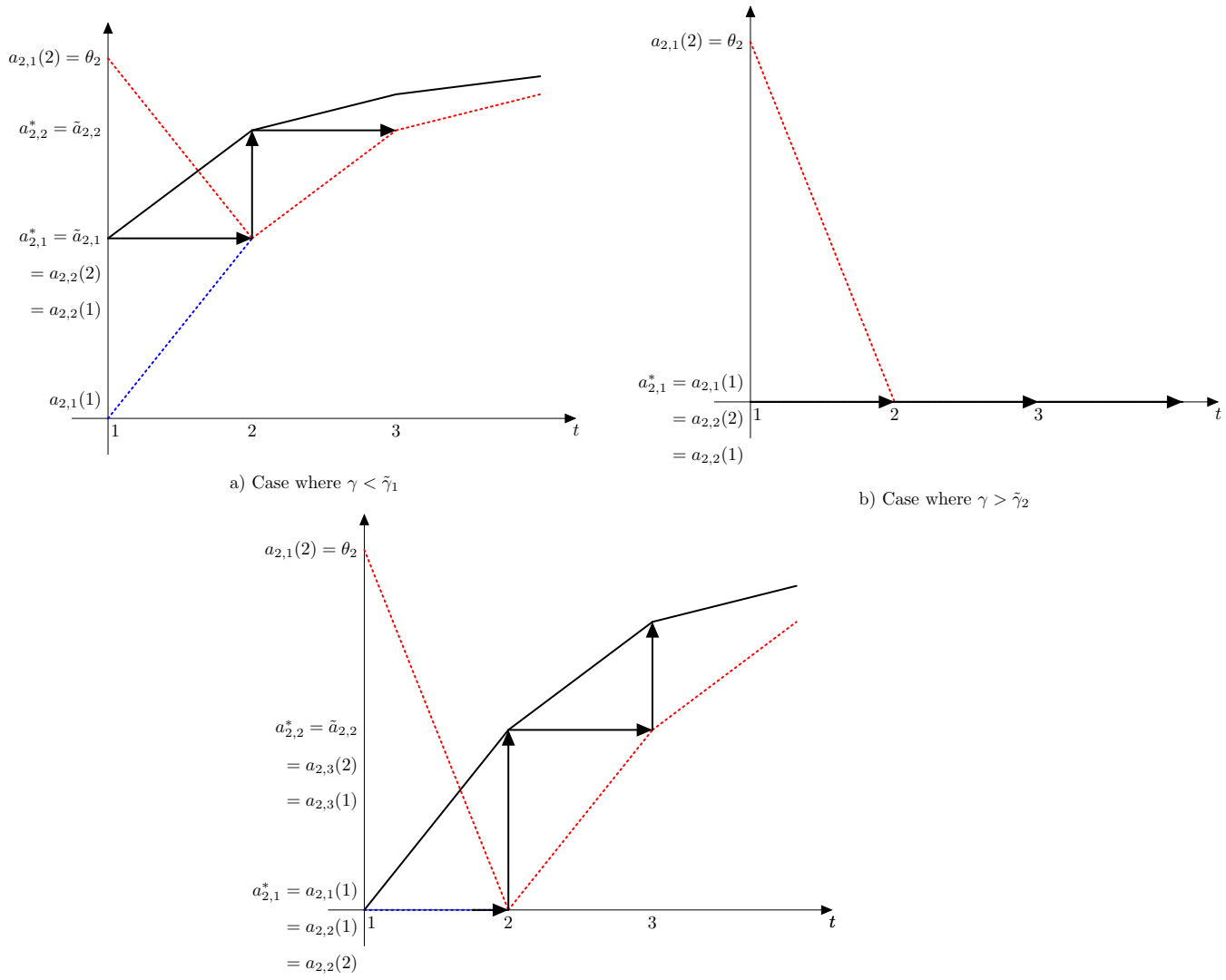
Theorem 4 *Denoting $\tilde{a}_{i,t}(\mathbf{r_t})$ the contribution effort that maximizes (2) in period t, there exist two thresholds $\tilde{\gamma}_1 < \tilde{\gamma}_2$ such that*

- *If $\gamma < \tilde{\gamma}_1$, $a^*_{i,t}(\mathbf{r}_t) = \tilde{a}_{i,t}(\mathbf{r_t})$ for any $i \in \mathcal{N}$ and $p^*_{12,t}(\mathbf{r_t}) = 0$ in any period t. Social roles and actions reflect abilities in the long run, $a_{i,\infty}(j) = a^*_{i,\infty} = \theta_j$ for any $i, j \in \mathcal{N}$.*

- *If $\gamma > \tilde{\gamma}_2$, $a^*_{i,t}(\mathbf{r_t}) = \tilde{a}_{i,t}(\mathbf{r_t})$ if $s_i = 1$ and $a^*_{i,t}(\mathbf{r}_t) = 0$ if $s_i = 2$ and $p^*_{12,t}(\mathbf{r_t}) > 0$ in any period t. In the long run, $a_{i,\infty}(j) = a^*_{i,\infty} = \theta_j$ if $s_j = 1$, and $a_{i,\infty}(j) = a^*_{i,\infty} = 0$ otherwise for any $i \in \mathcal{N}$.*

- *If $\tilde{\gamma}_1 < \gamma < \tilde{\gamma}_2$, then $a^*_{i,0}(\mathbf{r}_t) = \tilde{a}_{i,t}(\mathbf{r_t})$ if $s_i = 1$ and $a^*_{i,0}(\mathbf{r}_t) = 0$ otherwise and $p^*_{12,0}(\mathbf{r_t}) > 0$. For any period $t > 1$, $a^*_{i,t}(\mathbf{r}_t) = \tilde{a}_{i,t}(\mathbf{r_t})$ for any $i \in \mathcal{N}$ and $p^*_{12,t}(\mathbf{r_t}) = 0$. Social roles and actions reflect abilities in the long run, $a_{i,\infty}(j) = a^*_{i,\infty} = \theta_j$ for any $i, j \in \mathcal{N}$.*

The joint evolution of social roles and economic behaviors outlined in Theorem 4 is represented in the diagrams of Figure 1. The joint dynamics of social roles and economic behaviors displays two steady states. In the first steady state represented by panels a) and b), social roles and economic actions converge to the ability distribution. Each agent is socially perceived as able to contribute at a level that reflects his ability. In particular, type 2 agents do not face the threat of being punished, given that their social role is to contribute effort $\theta_2 > 0$ to the common pool. In the second steady state, represented by panel c), only type 1 agents contribute to the common pool. Everyone believes that the



a) Case where $\gamma < \tilde{\gamma}_1$

b) Case where $\gamma > \tilde{\gamma}_2$

c) Case where $\tilde{\gamma}_1 < \gamma < \tilde{\gamma}_2$

Figure 1: Equilibrium Dynamics

Note: the red dotted line represents $a_{2,t}(2)$, the blue dotted line represents $a_{2,t}(1)$ and the black line represents $a^*_{t,2}$.

social role of type 1 agents alone is to contribute, while the social role of type 2 agents is not to contribute. This belief about the social role of type 2 agents is sustained by a threat of punishment from type 1 agents.

Complementarity between economic behaviors and social roles characterizes the dynamics. Consider first panel a), which depicts a case where type 2 agents are not punished in equilibrium when they contribute. The type 2 agents initially contribute a certain level to the common pool and everyone revises their beliefs accordingly ($a^*_{2,1} = \tilde{a}_{2,1} = a_{2,2}(2) = a_{2,2}(1)$). In the next period, the type 2 agents contribute more, beliefs regarding their social role change again to reflect their higher contribution, and so forth until both the type 2 agents' contribution and their social role converge to their ability θ_2.

The same reasoning holds when the key complementarity between economic behaviors and social roles works in the opposite direction. Panel b) represents a case where type 2 agents are initially punished. Their initial contribution being equal to zero, everyone perceives these agents as non-contributors at the end of the first period ($a^*_{2,1} = a_{2,1}(1) = a_{2,2}(2) = a_{2,2}(1)$ with $a_{2,1}(1) = 0$). The type 1 agents will keep punishing the type 2 agents for seeking to contribute to the common pool. As a result, the long-run equilibrium is such that only type 1 agents contribute and are socially perceived as contributors.

Whether the steady state of panels a) or the steady state of panels b) and c) is reached hinges on the magnitude of γ. There are three cases to consider, as outlined in Theorem 4. In the first case, as represented in panel a), γ is low (i.e. $\gamma < \tilde{\gamma}_1$). Type 2 agents are not punished by type 1 agents for contributing in the first round of the game. As a result, beliefs on social roles change to reflect the ability of type 2 agents to contribute. Over time, this triggers a virtuous cycle between higher economic contributions from type 2 agents and an evolving perception of their social role.

In the second case, as represented in panel b), γ is high (i.e. $\gamma > \tilde{\gamma}_2$). Type 2 agents face a credible threat of punishment in period 1 if they deviate from the social role assigned to them by type 1 agents. Hence, they decide not to contribute in period 1. As a result, all the type 2 agents perceive their social role as not to contribute. Since γ is high, in period 2 type 2 agents still face punishment if they deviate from their social role. They abstain from contributing and the game reaches a steady state where only type 1 agents contribute and are socially perceived as sole contributors.

The third case is represented in panel c) and arises when γ takes intermediate values (i.e. $\tilde{\gamma}_1 < \gamma < \tilde{\gamma}_2$). Type 2 agents face a credible threat of punishment if they contribute in the first round of the game. Hence, they are deterred from contributing in the first period,

and this is reflected in changes to their beliefs regarding their own social role. They still have an incentive to contribute in the second round, albeit to a lesser extent than before. In the second period, the optimal contribution of the type 2 agents is low, so they are not threatened with punishment. As a result, their social role may be perceived slightly differently, now reflecting the fact that they can contribute too, albeit weakly. Type 2 agents can gradually increase their contribution and beliefs on their social role will follow suit, until an equilibrium is reached where both their contribution and their social role reflect their ability θ_2.

This last case illustrates a key result of this analysis. When an intermediate level of importance is ascribed to others' abiding by their social role, only large deviations from social roles are punished. Small deviations are not punished and arise endogenously after the first round of the game.

4 Stylized patterns

Gender roles. Across societies, people hold vastly differing beliefs on the appropriate social role of women. Alesina, Giuliano and Nunn (2013) trace these cultural differences back to gender roles in the traditional organization of agricultural production. Societies that traditionally practised plough agriculture developed a gender-based division of labor, with men tending to work in the fields and women active within the home. This division of labor generated preferences regarding the appropriate role of women in society.

My model describes how differences in one key parameter, γ, may have affected the evolution of gender roles and the economic trajectories of different societies. In one equilibrium, which is reached when γ is high, women remain inactive and the belief that women's social role is not to contribute is widespread. This might apply to societies that traditionally practised plough agriculture, given that gender roles were critical to the prosperity of these societies. In the context of the model, the belief that women's social role is not to contribute to the economy is strengthened by a threat of discrimination, domestic violence or harassment that women in these societies may still face. In the other equilibrium of the model, which is reached when γ is low, women gradually participate in economic production and beliefs on their social role are adjusted accordingly. This equilibrium might characterize the evolution of societies where economic production did not entail a gender-based division of labor.

Over the last century, developed economies witnessed a vast increase in female labor force participation (LFP). As demonstrated by Fernández (2013), this was accompanied by striking changes in social attitudes. The literature has proposed two closely related explanations for the joint evolution of social attitudes and female labor force participation. The first is that women were able to learn their own cost of working by observing the female labor supply in previous generations (Fernández (2013)). For example, in states where the mobilization rate was greater during World War II, there were more working women in 1950 (Acemoglu, Autor and Lyle (2004)). The next generation of women living in these states may have learned more about their own cost of working and therefore increased their labor supply.[13] The second explanation is rooted in the evolution of male attitudes toward female labor force participation. As hypothesized by Fernández, Fogli and Olivetti (2004), the increase in women's involvement in the formal labor market may have been driven by the increased number of men growing up with a family model where mothers work. The authors find that the probability of a man's wife working is significantly correlated with whether his mother worked.

The model developed in this paper squares these two theories of the evolution of female labor force participation, describing how female LFP evolves in tandem with both men's and women's beliefs on women's social role. The dynamics are characterized by the reinforcement over time of women's economic participation and of women's social role as workers. The model shows that men's beliefs affect women's beliefs regarding their own social role, by either rejecting or encouraging their economic participation. Hence, the model combines the two previous hypotheses within one unifying framework.

Finally, many studies have demonstrated that female leaders or role models can be a powerful inspiration to other women. For example, Beaman et al. (2012) showed that reserving leadership positions for women erased the gender gap in adolescent educational attainment and led girls to spend less time on household chores. Similarly, female science teachers and professors were found to boost female students' academic achievements (Dee (2007), Hoffmann and Oreopoulos (2009)). These findings are all consistent with the key mechanism of the model. The economic decisions made in one generation enable agents living in the next to adapt their beliefs on gender roles. As more women reach leadership positions or choose occupations and school curricula in traditionally male-dominated fields, both women and men can adapt their beliefs on the prevailing gender roles.

[13] Fernández, Fogli and Olivetti (2004) provide evidence in favor of this mechanism.

The Legacy of Slavery in the American South. Political and racial attitudes vary significantly across areas in the American South. As demonstrated by Acharya, Blackwell and Sen (2016), these disparities are partly rooted in the prevalence of slavery 150 years ago.[14] Similarly, several studies established a negative relationship between various measures of economic development and slavery in the United States (Mitchener and McLean (2003), Nunn (2008) and Lagerlöf (2006)).

The model describes how differences in parameter γ, which corresponds to how important it is to White workers that Black workers remain inferior in status and exploited, may have affected both the evolution of racial attitudes and economic outcomes. In one equilibrium, which is reached when γ is high, racist norms are widespread and Black workers are socially excluded. This equilibrium might reflect areas where factor endowments resulted in a more intensive use of slave labor in the antebellum South. Large-scale plantations necessitated more slave labor, which may have generated beliefs in White populations about the inferior status of Black workers and their social role as exploited labor. In the postbellum South, these racist norms were enforced through various means, including a system of racist laws and targeted violence against Blacks (Woodward (2002 [1955])).

According to the model, the negative relationship between various measures of economic development and slavery is explained by the widespread racism that constrains Black workers in their economic decisions. Importantly, the model also predicts that in this equilibrium, Blacks internalize beliefs on their own social role that sustain their social exclusion. These internalized beliefs may include a relationship between race, poverty or academic achievement. This prediction is consistent with the "culture of poverty" paradigm developed in sociology (Hannerz (1969), Lewis (1966), Riessman (1962) or Anderson (1990)). One manifestation of this culture might be the phenomenon commonly referred to as 'acting White', where in school, Black students may face costs for investing in behaviors more conducive to academic success, seen as characteristic of White students (Austen-Smith and Fryer (2005)). The model predicts that these patterns of beliefs, which help perpetuate the racial divide, might be a legacy of slavery, although this has not been investigated in the literature so far.

The other equilibrium of the model is reached when it is less important to White workers that Black workers remain inferior in status, i.e. when γ is low. In this equilibrium, Black

[14] Grosjean, Masera and Yousaf (2021) similarly finds that, in areas with a stronger history of slavery, a police officer is significantly more likely to stop a Black driver after a Trump rally during the 2015-2016 campaign . By implicitly associating violence with Blacks, Trump's speeches trigger deep-rooted stereotypes.

workers participate more in economic production, there is less social exclusion, and racist beliefs are weaker. The participation of Black workers in economic production weakens the racist beliefs of White workers and enables Black workers to revise their own beliefs regarding their social role. Black workers can see themselves as participating in economic production and having the same rights as White workers. This equilibrium might apply more to areas that relied less on slavery for their economic production in the antebellum south.

5 Welfare Analysis and Policy Implications

5.1 Welfare

We restrict our attention to the case where all agents have the same ability, $\theta_1 = \theta_2 = \theta$. I denote W_∞ the long-run social welfare of agents. From Theorem 4, it is only when $\gamma < \tilde{\gamma}_2$ that long-run contributions reflect abilities. Otherwise, the long-run contribution of type 2 agents remains equal to zero. Hence,

$$W_\infty = \begin{cases} n(n - \frac{1}{2})\theta^2 \text{ if } \gamma < \tilde{\gamma}_2 \text{ and} \\ (n - n_2)(n - \frac{1}{2})\theta^2 \text{ otherwise.} \end{cases} \tag{4}$$

Theorem 5 *The equilibrium reached when $\gamma_2 > \tilde{\gamma}_2$ is Pareto dominated by the equilibrium reached when $\gamma_2 < \tilde{\gamma}_2$.*

When $\gamma > \tilde{\gamma}_2$, only type 1 agents contribute and are socially perceived as sole contributors. From Theorem 5, both type 1 and type 2 agents would have been better off in the equilibrium where all agents contribute. Indeed, in the long-run equilibrium where $\gamma > \tilde{\gamma}_2$, type 1 agents do not benefit from the contributions of type 2 agents. Type 2 agents reach lower utility levels too, as their contribution is lower than what they perceive as optimal.

Given that the long-run equilibrium in the case where $\gamma_2 > \tilde{\gamma}_2$ is Pareto dominated, there is scope for public interventions that decrease the likelihood of reaching the inefficient equilibrium. In the next subsection, I establish two main policy implications.

5.2 Quotas and other forms of positive discrimination

More often than not, firms, organizations or societies contain individuals of different genders, cultures, races, religions or sexual identities. Hence, a key policy question is what conditions are required for heterogeneous groups to reach optimal production levels.

Theorem 6 *There exists a threshold $\tilde{n}_2$ such that if $n_2 < \tilde{n}_2$, long-run social welfare is decreasing in n_2. If $n_2 \geq \tilde{n}_2$, long-run social welfare is maximum and remains constant.*

This result shows that effectively promoting diversity can be a powerful way to eliminate a punishment threat weighing on type 2 agents when they do not conform to their assigned social role. Indeed, when there is a large enough number of type 2 agents in the production group, it becomes too costly for type 1 agents to punish them, since that would involve forgoing a significant share of the group production. As a result, punishment disappears and beliefs regarding the social roles of type 2 agents can change.

This result strongly supports policies such as quotas and other forms of positive discrimination that increase the participation of women and ethnic or religious minorities. Quotas have been shown to affect beliefs on gender roles (Beaman et al. (2009)). Similarly, Bastian (2020) shows that the 1975 introduction of Earned Income Tax Credit in the United States increased maternal employment and led to increased approval of working women. The analysis also supports policies that aim at reducing racial inequalities by creating heterogeneous neighborhoods (Chetty and Hendren (2018)). Importantly, it demonstrates that quotas and other forms of positive discrimination not only decrease discrimination in the short run, but also contribute to changing beliefs on social roles. Such positive discrimination can therefore have key long-run effects on overall economic production and social welfare.

5.3 Laws and other constraints reducing punishment

We consider a simple extension of the model where type 2 agents can now whistle-blow regarding the punishment they undergo. Formally, when an agent of type 1 punishes an agent of type 2 in period t, he is reported and has to pay an additional cost $q > 0$. The cost parameter q reflects the legal constraints on punishers as well as the social stigma that they may face when the punishment is made known. For example, harassment and various forms of discrimination against women in the workplace are illegal in many countries.

Theorem 7 *When $\gamma > \tilde{\gamma}_2$, there exist a threshold $\tilde{q} > 0$ and a threshold τ such that if $q > \tilde{q}$ for at least τ periods of time, then social roles and actions reflect abilities in the long run, $a_{i,\infty}(j) = a^*_{i,\infty} = \theta_j$ for any $i, j \in \mathcal{N}$.*

When $\gamma > \tilde{\gamma}_2$, from Theorem 4, the long-run equilibrium should be such that only type 1 agents contribute to the common pool. However, if society imposes for a finite period τ a sufficiently high cost on type 1 agents for punishing type 2 agents (i.e. $q > \tilde{q}$), then type 2 agents will be able to produce without facing punishment. The perception of their social role will gradually change. After τ periods of time, the social roles will be such that type 1 agents believe that it is also the social role of type 2 agents to contribute to the common pool. The Pareto inefficient equilibrium outlined in Theorem 4 will not be reached.

This result shows that both legal constraints and short-run costs imposed on the perpetrators of domestic violence, harassment or racial discrimination can be instrumental in changing long-run beliefs on social roles in society at large. As an illustration, White Americans' attitudes toward racial desegregation and toward Black Americans became more progressive around the time of the Civil Rights and Black Power movements in the 1960s and early 1970s. More recently, in 2013, the #BlackLivesMatter movement, aimed at denouncing instances of police violence against Blacks, raised the cost of racial violence for police officers. From the viewpoint of the model, the movement may have changed beliefs on the social role of Blacks by temporarily reducing their social exclusion. This prediction accords well with the recent study of Sawyer and Gampa (2018), who find that events associated with the #BlackLivesMatter movement are associated with less pro-White attitudes, as measured through implicit association tests in a large sample across the United States. The #MeToo movement also temporarily increased the cost of sexual harassment and sexual abuse of women by making allegations public. This movement may have changed beliefs on gender roles by enabling women to participate more in economic activities of their choosing.

5.4 Costly Whistle-Blowing

Whistle-blowing can be a costly strategy for victims of discrimination or stigma, as they might face reprisals. Although whistle-blowers are protected by law in the United States and many companies have whistle-blower protection policies, employees often face threats of retaliation (Rehg et al. (2008)).

I introduce a whistle-blowing cost $\epsilon > 0$ for the victims of punishment. Once an agent of type 1 is reported, he has to pay $q > 0$ when he punishes an agent of type 2. As in Section 5.3, $q > 0$ reflects both the legal constraints on punishers and the social stigma they may face when the punishment is made common knowledge.

Theorem 8 *When $\gamma > \tilde{\gamma_2}$, there exist a threshold $\tilde{q}(\epsilon) > 0$ and a threshold $\tau(\epsilon)$ both increasing with ϵ such that if $q > \tilde{q}(\epsilon)$ for at least $\tau(\epsilon)$ periods of time, then social roles and actions reflect abilities in the long run, $a_{i,\infty}(j) = a^*_{i,\infty} = \theta_j$ for any $i, j \in \mathcal{N}$.*

This result is a generalization of Theorem 7 to the case where $\epsilon > 0$. Since whistle-blowing is costly for type 2 agents, it takes less punishment to muzzle them. Hence, to eliminate the punishment threat, society must impose an even higher cost $q(\epsilon)$ on punishers than in the case where $\epsilon = 0$.

Importantly, when whistle-blowing is costly, it takes longer for social roles to change, i.e $\tau(\epsilon)$ increases with ϵ, because punishment is cheap for type 1 agents. Theorem 8 shows that when the indirect costs associated with whistle-blowing are high, legal constraints are less effective in preventing harassment and other forms of discrimination. Hence, Theorem 8 supports both continuing legal constraints and protection policies within organizations that protect whistle-blowers from retaliation.

6 Other game settings

The model extends to two other standard game settings.

Self-nomination games. First, the model can be extended to game settings where an individual can self-nominate to perform a task that benefits everyone in the group. The monetary payoff of everyone in the group will then be equal to the efficient contribution of the self-nominated individual. This type of game corresponds to the following functional form for the monetary payoff $x(a_i, \mathbf{a_{-i}})$:

$$x(a_i, \mathbf{a_{-i}}) = \max_{j \in \mathcal{N}} \theta_j a_j$$

if $j \neq k$, with $k = \arg\max_{j \in \mathcal{N}} \theta_j a_j$, and

$$x(a_k, \mathbf{a_{-k}}) = \theta_k a_k - \frac{a_k^2}{2}.$$

This monetary payoff is in line with the experimental design of Coffman (2014). Agents choose their willingness to contribute answers in a given group task, e.g. their willingness to self-nominate to answer a mathematical question. The group answer that is then selected represents the answer making the greatest contribution. In this case, a_i could indicate the willingness to self-nominate of individual i.

All the results of this paper extend to this type of game, including the welfare analysis. Consider for example the case where type 1 agents initially believe that the social role of type 2 agents is not to self-nominate (i.e. $a_{2,1}(1) = 0$) and this belief is of great importance to them (i.e. γ is high). The revised version of Theorem 4 then implies that type 2 agents will not self-nominate even if they are very capable of performing the task at hand. As a result, everyone will tend to perceive that the social role of type 2 agents is not to self-nominate. In the long run, every agent could have been made better off if type 2 agents were able to self-nominate.

The results in this game shed light on several relevant dimensions of the complex impact of gender roles on aggregate behavior. In particular, they explain why men are more willing to self-nominate in many fields that are traditionally perceived as masculine (i.e, STEM fields, business). Moreover, they also explain why women internalize the belief that they should not self-nominate in these fields, as shown by Coffman (2014) in experimental settings.[15]

Market games. The model can also be extended to game settings where there is an exchange between two or more individuals. For simplicity, consider that there are only two agents, a buyer and a seller. The seller seeks to sell a unique good that is valued by the buyer. For example, the seller could be a worker supplying labor, a business entrepreneur or a long-distance merchant in a more historical context.

In each period, seller s decides whether or not to sell the good to the buyer. Hence, $a_s = 1$ if the seller decides to be active and sell her good to the buyer and $a_s = 0$ otherwise. If the seller decides to sell the good to the buyer, she incurs a sunk production cost $c > 0$. For example, c could be the cost of job-hunting for a worker supplying her labor. The seller also sets price r of the good she is selling. If the seller decides to be active, buyer j decides whether or not to buy the good. Hence, $a_b = 1$ if the buyer decides to buy the good, and $a_b = 0$ otherwise. The value of the good is $V_b > 0$. I assume that the buyer can

[15] Similarly, Cooper and Kagel (2016) finds that in teams, women are much less likely to advocate strategic play than men.

borrow against future income, for simplicity, and V_b is drawn from a uniform distribution on segment $[0, 1]$.

I assume that the buyer initially believes that the social role of the seller is not to be active, i.e. $a_s(b) = 0$. The seller believes that she should be active, and $a_s(s) = 1$. One example of the many situations of this kind is a male employer believing that the social role of a female job candidate is not to be on the labor market. To further simplify, I assume that $\alpha_{bb} = \alpha_{bs} = 0$, so the buyer does not have social image concerns when he decides to buy the good. Finally, I assume that $\gamma_{bs} \equiv \gamma > 0$, while $\gamma_{sb} = 0$. The seller's action matters to the buyer, while the seller does not care what the buyer does.

Under these conditions, the utility of the buyer writes:

$$u_b(a_b, a_s, q) = a_b a_s (V_b - r) - p_{bs} - \frac{\gamma}{2} a_s^2,$$

as he obtains a monetary payoff $V_b - r$ only when he buys the good (i.e. when $a_b a_s = 1$). The seller is uncertain about the value of the good to the buyer. Hence, her expected utility is:

$$\mathbb{E}\, u_s(a_s, a_b, q) = a_s(-c + \pi(r)r) - p_{bs} - \frac{\alpha_{ss}}{2}(1 - a_s)^2 - \frac{\alpha_{sb}}{2} a_s^2$$

with $\pi(r)$ the probability that the exchange occurs at price r.

Theorem 9 *There exists a threshold $\tilde{\gamma}$ such that*

- *If $\gamma < \tilde{\gamma}$ and $c < 1/4 - (\alpha_{bs} - \alpha_{ss})/2$, $p_{bs,\infty} = 0$ in the long run and $a_{b,\infty} = a_{b,\infty}(k) = 1$ for any $k \in \{b, s\}$, while in any period t, $a_{s,t} = 1$ if $w_{b,t} > 1/2$ and $a_{s,t} = 0$ otherwise.*
- *If $\gamma \geq \tilde{\gamma}$ or $c \geq 1/4 - (\alpha_{bs} - \alpha_{ss})/2$, $p_{bs,\infty} = \max(0, 1/4 - c - (\alpha_{bs} + \alpha_{bb})/2)$ in the long run and $a_{b,\infty} = a_{b,\infty}(k) = 0$ for any $k \in \{b, s\}$ and $a_{s,\infty} = 0$.*

Theorem 9 is a generalization of Theorem 4 to a simple market game, as outlined above. If the buyer believes that the social role of the seller is to remain inactive and γ is high ($\gamma \geq \tilde{\gamma}$), then the buyer will remain inactive in the long run. Both the buyer and the seller will internalize beliefs on their social roles such that the buyer will not be active on the market. By contrast, if γ is sufficiently low ($\gamma < \tilde{\gamma}$), then the buyer will be active (provided that her utility from doing so is positive, i.e. $c < 1/4 - (\alpha_{bs} - \alpha_{ss})/2$) and the seller will revise his internalized beliefs on the social role of the buyer. The other results of the paper extend as well. In particular, both the seller and the buyer are worse off in the long-run

equilibrium where the seller is inactive, as both would have benefited from an exchange occurring.

This simple market game also describes important cases of social roles in application and describes how they can generate market failures. It provides insights into how internalized beliefs on gender roles create the labor market discrimination against women widely documented in the literature. It can obviously apply to other forms of labor market discrimination against Black workers, minorities, religious or ethnic groups, also the subject of a vivid empirical literature.[16]

Finally, this game explains how social roles structure the functioning of exchange markets. Trade is often codified and embedded in social roles. For example, in his extensive study of primitive economies, (Sahlins, 2017 [1974], p. 168) argues that "a material transaction is usually a momentary episode in a continuous social relation. The social relation exerts governance: the flow of goods is constrained by, is part of, a status etiquette". Historically, religions have also structured exchange markets, not only by providing legal constraints but also by codifying exchange with social and symbolic meanings. This may have led to internalized norms that make individuals more inclined to trade with people sharing their faith (Chaudhuri (1985)). Finally, even in industrialized economies, the logic of exchange remains marked by social codification, and business cultures vary across countries.[17]

7 Discussion

In this paper, I introduced a dynamic utility model of the interplay between economic actions and social roles. I modeled both how economic actions are embedded in social roles and how social roles reciprocally feed back into internalized preferences and affect economic outcomes.

[16] See, among others, Bertrand and Mullainathan (2004) on labor market discrimination against women or Adida, Laitin and Valfort (2016) on labor market discrimination against Muslims. This dynamic model shows that taste-based discrimination and statistical discrimination are in fact closely related, although they are often distinguished in the economic literature. Internalized beliefs on social roles not only explain why agents in hiring positions will tend to discriminate against members of distinctive groups that they dislike (taste-based discrimination). They also explain why agents belonging to the group discriminated against will internalize beliefs on their own social role that tend to reinforce the discrimination they experience. Hence, internalized beliefs among victims of discrimination can potentially create a variety of endogenous behaviors that provide a rational basis for statistical discrimination.

[17] See, for example, Hofstede (1994).

I demonstrated that this analysis generates rich behavioral dynamics explaining a wide range of empirical and experimental regularities, while at the same time providing an interpretation grid for the historical evolution of social roles and development outcomes. I discussed in particular the evolution of gender roles and the persistence of racial divides.

I also found that the joint evolution of social roles and economic outcomes has key welfare implications. Across standard game settings, when some individuals oblige others to conform to their beliefs on social roles, a Pareto dominated equilibrium is reached in the long run. I find that policies or social movements that give oppressed groups a *window of opportunity* to realize their economic potential can challenge inefficient beliefs on social roles.

This model has several important limitations. First, social roles are multidimensional, and accounting for this in future research might explain a new set of empirical findings on strategic identification with different social roles. Indeed, the internal design of beliefs on social roles might be affected by which social roles agents choose to adopt before making economic decisions.[18] A multidimensional extension of this work could also help explain how social roles that are not acted out can nevertheless persist (Greif and Tadelis (2010)). Second, this paper considers a fundamental interplay between prevailing economic outcomes and the evolution of internalized beliefs on social roles. Yet cultures may react endogenously and promote worldviews that fight the pressure imposed by prevailing economic conditions on the evolution of preferences. For example, the Passover ritual consists in every year reading and asking questions about the exodus of the Jews from Egypt. Key to the ritual is the affirmation of emancipation, both from the prevailing economic order (slavery) and from the internalized beliefs that sustain it.

References

Abeler, Johannes, Daniele Nosenzo and Collin Raymond. 2019. "Preferences for Truth-Telling." *Econometrica* 87(4):1115–1153.

Acemoglu, Daron, David H. Autor and David Lyle. 2004. "Women, War, and Wages: The Effect of Female Labor Supply on the Wage Structure at Midcentury." *Journal of Political Economy* 112(3):497–551.

[18] There is an emerging literature on identity choice. See, for example, Atkin, Colson-Sihra and Shayo (2021) on ethnic or religious identity choices and Seror and Ticku (2020) on sexual identity choices.

Acharya, Avidit, Matthew Blackwell and Maya Sen. 2016. "The Political Legacy of American Slavery." *The Journal of Politics* 78:000–000.

Adida, Claire L., David D. Laitin and Marie-Anne Valfort. 2016. ""One Muslim is Enough!" Evidence from a Field Experiment in France." *Annals of Economics and Statistics* (121/122):121–160.

Akerlof, George A. and Rachel E. Kranton. 2000. "Economics and Identity*." *The Quarterly Journal of Economics* 115(3):715–753.

Akerlof, George and William Dickens. 1982. "The Economic Consequences of Cognitive Dissonance." *American Economic Review* 72:307–19.

Akerlof, Robert. 2017. "Value Formation: The Role of Esteem." *Games and Economic Behavior* 102(C):1–19.

Alesina, Alberto and Paola Giuliano. 2015. "Culture and Institutions." *Journal of Economic Literature* 53(4):898–944.

Alesina, Alberto, Paola Giuliano and Nathan Nunn. 2013. "On the Origins of Gender Roles: Women and the Plough." *Quarterly Journal of Economics* 128(2):469–530.

Alger, Ingela and Jörgen W. Weibull. 2013. "Homo Moralis—Preference Evolution Under Incomplete Information and Assortative Matching." *Econometrica* 81(6):2269–2302.

Anderson, E. 1990. *Streetwise: Race, Class, and Change in an Urban Community.* University of Chicago Press.

Atkin, David, Eve Colson-Sihra and Moses Shayo. 2021. "How Do We Choose Our Identity? A Revealed Preference Approach Using Food Consumption." *Journal of Political Economy* 129(4):1193–1251.

Austen-Smith, David and Roland G. Fryer. 2005. "An Economic Analysis of "Acting White"." *The Quarterly Journal of Economics* 120(2):551–583.

Bastian, Jacob. 2020. "The Rise of Working Mothers and the 1975 Earned Income Tax Credit." *American Economic Journal: Economic Policy* 12(3):44–75.

Beaman, Lori, Esther Duflo, Rohini Pande and Petia Topalova. 2012. "Female Leadership Raises Aspirations and Educational Attainment for Girls: A Policy Experiment in India." *Science* 335(6068):582–586.

Beaman, Lori, Raghabendra Chattopadhyay, Esther Duflo, Rohini Pande and Petia Topalova. 2009. "Powerful Women: Does Exposure Reduce Bias?" *The Quarterly Journal of Economics* 124(4):1497–1540.

Bénabou, Roland and Jean Tirole. 2002. "Self-Confidence and Personal Motivation*." *The Quarterly Journal of Economics* 117(3):871–915.

Bénabou, Roland and Jean Tirole. 2003. "Intrinsic and Extrinsic Motivation." *Review of Economic Studies* 70(3):489–520.

Bénabou, Roland and Jean Tirole. 2004. "Willpower and Personal Rules." *Journal of Political Economy* 112(4):848–886.

Bénabou, Roland and Jean Tirole. 2011*a*. " Identity, Morals, and Taboos: Beliefs as Assets *." *The Quarterly Journal of Economics* 126(2):805–855.

Bénabou, Roland and Jean Tirole. 2011*b*. "Identity, morals, and taboos: Beliefs as assets." *Quarterly Journal of Economics* 126(2):pp. 805–855.

Bensoussan, G. 2012. *Juifs en pays arabes: le grand déracinement, 1850-1975*. Histoires d'aujourd'hui Tallandier.

Bernheim, B. Douglas, Luca Braghieri, Alejandro Martínez-Marquina and David Zuckerman. 2021. "A Theory of Chosen Preferences." *American Economic Review* 111(2):720–54.

Bertrand, Marianne. 2011. Chapter 17 - New Perspectives on Gender. Vol. 4 of *Handbook of Labor Economics* Elsevier pp. 1543–1590.

Bertrand, Marianne. 2020. "Gender in the Twenty-First Century." *AEA Papers and Proceedings* 110:1–24.

Bertrand, Marianne, Emir Kamenica and Jessica Pan. 2015. " Gender Identity and Relative Income within Households." *The Quarterly Journal of Economics* 130(2):571–614.

Bertrand, Marianne and Sendhil Mullainathan. 2004. "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *The American Economic Review* 94(4):991–1013.

Besley, Timothy. 2020. "State Capacity, Reciprocity, and the Social Contract." *Econometrica* 88(4):1307–1335.

Besley, Timothy and Torsten Persson. 2019. "Democratic Values and Institutions." *American Economic Review: Insights* 1(1):59–76.

Bisin, Alberto, Jared Rubin, Avner Seror and Thierry Verdier. 2021. Culture, Institutions & the Long Divergence. NBER Working Papers 28488 National Bureau of Economic Research, Inc.

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. 2011. Chapter 9 - The Economics of Cultural Transmission and Socialization. Vol. 1 of *Handbook of Social Economics* North-Holland pp. 339 – 416.

Bisin, Alberto and Thierry Verdier. 2017. "On the Joint Evolution of Culture and Institutions." NBER Working Paper No. 23375.

Bohnet, Iris, Alexandra van Geen and Max Bazerman. 2016. "When Performance Trumps Gender Bias: Joint vs. Separate Evaluation." *Management Science* 62(5):1225–1234.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli and Andrei Shleifer. 2016. "Stereotypes." *Quarterly Journal of Economics* 131(4):1753–1794.

Bowles, Samuel. 1998. "Endogenous Preferences: The Cultural Consequences of Markets and Other Economic Institutions." *Journal of Economic Literature* 36(1):75–111.

Bramoullé, Yann and Rachel Kranton. 2007. "Public goods in networks." *Journal of Economic Theory* 135(1):478–494.

Bursztyn, Leonardo, Alessandra L. González and David Yanagizawa-Drott. 2020. "Misperceived Social Norms: Women Working Outside the Home in Saudi Arabia." *American Economic Review* 110(10):2997–3029.

Bursztyn, Leonardo, Georgy Egorov and Stefano Fiorin. 2020. "From Extreme to Mainstream: The Erosion of Social Norms." *American Economic Review* 110(11):3522–48.

Bursztyn, Leonardo, Thomas Fujiwara and Amanda Pallais. 2017. "'Acting Wife': Marriage Market Incentives and Labor Market Investments." *American Economic Review* 107(11):3288–3319.

Chaudhuri, K.N. 1985. *Trade and Civilisation in the Indian Ocean: An Economic History from the Rise of Islam to 1750*. Cambridge paperback library Cambridge University Press.

Chetty, Raj and Nathaniel Hendren. 2018. "The Impacts of Neighborhoods on Intergenerational Mobility I: Childhood Exposure Effects*." *The Quarterly Journal of Economics* 133(3):1107–1162.

Coffman, Katherine Baldiga. 2014. "Evidence on Self-Stereotyping and the Contribution of Ideas." *The Quarterly Journal of Economics* 129(4):1625–1660.

Cooper, David J. and John H. Kagel. 2016. "A failure to communicate: an experimental investigation of the effects of advice on strategic play." *European Economic Review* 82:24–45.

Croson, Rachel and Uri Gneezy. 2009. "Gender Differences in Preferences." *Journal of Economic Literature* 47(2):448–74.

de Foucauld, C. 1998. *Reconnaissance au Maroc: 1883-1884*. Introuvables (Harmattan (Firm))) L'Harmattan.

Dee, Thomas S. 2007. "Teachers and the Gender Gaps in Student Achievement." *Journal of Human Resources* 42(3).

Eckel, Catherine and Philip Grossman. 2008. "Men, Women and Risk Aversion: Experimental Evidence." *Handbook of experimental economics results* 1.

Fehr, Ernst and Klaus M. Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation." *The Quarterly Journal of Economics* 114(3):817–868.

Fernández, Raquel. 2013. "Cultural Change as Learning: The Evolution of Female Labor Force Participation over a Century." *American Economic Review* 103(1):472–500.

Fernández, Raquel, Alessandra Fogli and Claudia Olivetti. 2004. "Mothers and Sons: Preference Formation and Female Labor Force Dynamics*." *The Quarterly Journal of Economics* 119(4):1249–1299.

Gilardi, Fabrizio. 2015. "The Temporary Importance of Role Models for Women's Political Representation." *American Journal of Political Science* 59(4):957–970.

Granovetter, Mark. 1985. "Economic Action and Social Structure: The Problem of Embeddedness." *American Journal of Sociology* 91(3):481–510.

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

Grosjean, Pauline, Federico Masera and Hasin Yousaf. 2021. Whistle the Racist Dogs: Political Campaigns and Police Stops. CEPR Discussion Papers 15691 C.E.P.R. Discussion Papers.

Hannerz, U. 1969. *Soulside: Inquiries Into Ghetto Culture and Community*. Almqvist & Wiksell (distr.).

Hoffmann, Florian and Philip Oreopoulos. 2009. "A Professor Like Me: The Influence of Instructor Gender on College Achievement." *Journal of Human Resources* 44(2).

Hofstede, Geert. 1994. "The business of international business is culture." *International Business Review* 3(1):1–14.

Inglehart, Ronald and Pippa Norris. 2003. *Rising tide : gender equality and cultural change around the world*. Cambridge, UK New York: Cambridge University Press.

Jayachandran, Seema. 2015. "The Roots of Gender Inequality in Developing Countries." *Annual Review of Economics* 7(1):63–88.

Johnson, N.D. and M. Koyama. 2019. *Persecution & Toleration: The Long Road to Religious Freedom*. Cambridge Studies in Economics, Choice, and Society Cambridge University Press.

Kahneman, Daniel, Peter P. Wakker and Rakesh Sarin. 1997. "Back to Bentham? Explorations of Experienced Utility." *The Quarterly Journal of Economics* 112(2):375–405.

Kuran, Timur and Scott Lustig. 2012. "Judicial Biases in Ottoman Istanbul: Islamic Justice and Its Compatibility with Modern Economic Life." *The Journal of Law Economics* 55(3):631–666.

Lagerlöf, Nils-Petter. 2006. Geography, institutions, and growth: the United States as a microcosm. Technical report.

Lewis, O. 1966. *La Vida: A Puerto Rican Family in the Culture of Poverty–San Juan and New York*. A Vintage giant Random House.

Mitchener, Kris and Ian McLean. 2003. "The Productivity of US States since 1880." *Journal of Economic Growth* 8.

Montgomery, James D. 1998. "Toward a Role-Theoretic Conception of Embeddedness." *American Journal of Sociology* 104(1):92–125.

Montgomery, James D. 2004. "The logic of role theory: Role conflict and stability of the self-concept." *The Journal of Mathematical Sociology* 29(1):33–71.

Niederle, Muriel and Lise Vesterlund. 2007. "Do Women Shy Away From Competition? Do Men Compete Too Much?*." *The Quarterly Journal of Economics* 122(3):1067–1101.

Niederle, Muriel and Lise Vesterlund. 2010. "Explaining the Gender Gap in Math Test Scores: The Role of Competition." *Journal of Economic Perspectives* 24(2):129–44.

Nunn, Nathan. 2008. *Slavery, Inequality, and Economic Development in the Americas: An Examination of the Engerman-Sokoloff Hypothesis*. Cambridge: Harvard University Press pp. 148–180.

Olivetti, Claudia and Barbara Petrongolo. 2016. "The Evolution of Gender Gaps in Industrialized Countries." *Annual Review of Economics* 8(1):405–434.

Oxoby, Robert J. 2003. "Attitudes and allocations: status, cognitive dissonance, and the manipulation of attitudes." *Journal of Economic Behavior Organization* 52(3):365–385.

Oxoby, Robert J. 2004. "Cognitive dissonance, status and growth of the underclass*." *The Economic Journal* 114(498):727–749.

Rabin, Matthew. 1994. "Cognitive dissonance and social change." *Journal of Economic Behavior Organization* 23(2):177–194.

Rehg, Michael T., Marcia P. Miceli, Janet P. Near and James R. Van Scotter. 2008. "Antecedents and Outcomes of Retaliation against Whistleblowers: Gender Differences and Power Relationships." *Organization Science* 19(2):221–240.

Reuben, Ernesto, Paola Sapienza and Luigi Zingales. 2014. "How stereotypes impair women's careers in science." *Proceedings of the National Academy of Sciences* 111(12):4403–4408.

Riessman, F. 1962. *The Culturally Deprived Child*. Harper.

Robson, Arthur J. and Larry Samuelson. 2011. The Evolutionary Foundations of Preferences. Vol. 1 of *Handbook of Social Economics* North-Holland pp. 221–310.

Sahlins, M. 2017 [1974]. *Stone Age Economics*. Routledge Classics Taylor & Francis.

Sambanis, Nicholas and Moses Shayo. 2013. "Social Identification and Ethnic Conflict." *American Political Science Review* 107(2):294–325.

Sawyer, Jeremy and Anup Gampa. 2018. "Implicit and Explicit Racial Attitudes Changed During Black Lives Matter." *Personality and Social Psychology Bulletin* 44(7):1039–1059. PMID: 29534647.

Seror, Avner and Rohit Ticku. 2020. Sexual Identity and Priesthood: Theory and Evidence. Technical report.

Shayo, Moses. 2009. "A Model of Social Identity with an Application to Political Economy: Nation, Class, and Redistribution." *The American Political Science Review* 103(2):147–174.

Shayo, Moses. forthcoming. "Social Identity and Economic Policy." *Annual Review of Economics* .

Stryker, Sheldon and Peter J. Burke. 2000. "The Past, Present, and Future of an Identity Theory." *Social Psychology Quarterly* 63(4):284–297.

Tabellini, Guido. 2008. "The Scope of Cooperation: Values and Incentives." *The Quarterly Journal of Economics* 123(3):905–950.

Woodward, C. Van. 2002 [1955]. *The Strange Career of Jim Crow*. Oxford University Press.

For Online Publication

Supplement to "Social Roles"

A Theory Appendix

A.1 Proof of Theorem 1

In the case where $s_i = 1$ for $i \in \mathcal{N}$, the derivative of (2) with respect to a_1 writes:

$$\frac{\partial u_1}{\partial a_1} = -a_1 + \theta_1 - n_1\alpha_{11} \mid a_1 - a_1(1) \mid -n_2\alpha_{12} \mid a_1 - a_1(2) \mid . \tag{A.1}$$

Given that $1 > \alpha_{11}n_1 + \alpha_{12}n_2$ by assumption, $\frac{\partial u_1}{\partial a_1}$ is decreasing in a_1 and continuous. I deduce that equation

$$\frac{\partial u_1}{\partial a_1} = 0 \tag{A.2}$$

admits a unique solution, which I denote $\tilde{a}_1$. A similar reasoning applies in the case where $s_i = 2$ for $i \in \mathcal{N}$ and I denote $\tilde{a}_2$ the unique solution of $\frac{\partial u_2}{\partial a_2} = 0$.

The individuals of type 2 cannot punish those of type 1. Hence, the optimal action chosen by an individual of type 1 is necessarily $\tilde{a}_1$, $a_1^* = \tilde{a}_1$.

The individuals of type 1 can punish those of type 2. We can now derive the optimal punishment strategy of the individuals of type 1.

The minimum punishment p_{12}^* is set by individuals of type 1 so that individuals of type 2 are made indifferent between choosing their optimal action $\tilde{a}_2$ and conforming to the social role assigned to them by type 1. Hence, if there is a punishment in equilibrium, then

$$p_{12}^* = u_2(a_1^*, \tilde{a}_2) - u_2(a_1^*, a_2(1)). \tag{A.3}$$

Such a punishment is incentive-compatible for the individuals of type 1 when

$$p_{12}^* < u_1(a_1^*, \tilde{a}_2) - u_1(a_1^*, a_2(1)). \tag{A.4}$$

We find that

$$p_{12}^* = (\tilde{a}_2 - a_2(1))\{n_2\theta_2 - 1/2(\tilde{a}_2 + a_2(1)) + n_2\alpha_{22}/2(2a_2(2) - a_2(1) - \tilde{a}_2)\} \tag{A.5}$$

and the incentive compatibility constraint rewrites

$$p_{12}^* < (\tilde{a}_2 - a_2(1))\{-n_2\theta_2 + n_2\gamma_{12}/2(\tilde{a}_2 - a_2(1))\}. \tag{A.6}$$

We deduce that in the case where $\tilde{a}_2 > a_2(1)$, the last inequality writes

$$n_2\theta_2 - 1/2(\tilde{a}_2 + a_2(1)) + n_2\alpha_{22}/2(2a_2(2) - a_2(1) - \tilde{a}_2) - 1/2\alpha_{21}n_1(\tilde{a}_2 - a_2(1)) < -n_2\theta_2 + n_2\gamma_{12}/2(\tilde{a}_2 - a_2(1)) \tag{A.7}$$

or

$$\gamma_{12} > \frac{2n_2\theta_2 - 1/2(\tilde{a}_2 + a_2(1)) - 1/2\alpha_{21}n_1(\tilde{a}_2 - a_2(1)) + n_2\alpha_{22}/2(2a_2(2) - a_2(1) - \tilde{a}_2)}{n_2/2(\tilde{a}_2 - a_2(1))}. \tag{A.8}$$

When $\tilde{a}_2 < a_2(1)$, it rewrites

$$n_2\theta_2 - 1/2(\tilde{a}_2 + a_2(1)) - 1/2\alpha_{21}n_1(\tilde{a}_2 - a_2(1)) + n_2\alpha_{22}/2(2a_2(2) - a_2(1) - \tilde{a}_2) > -n_2\theta_2 + n_2\gamma_{12}/2(\tilde{a}_2 - a_2(1)) \tag{A.9}$$

or

$$\gamma_{12} > \frac{-2n_2\theta_2 + 1/2(\tilde{a}_2 + a_2(1)) + 1/2\alpha_{21}n_1(\tilde{a}_2 - a_2(1)) - n_2\alpha_{22}/2(2a_2(2) - a_2(1) - \tilde{a}_2)}{n_2/2(a_2(1) - \tilde{a}_2)}. \tag{A.10}$$

In both cases, there exists a threshold value $\tilde{\gamma}$ such that if $\gamma_{12} > \tilde{\gamma}$, the punishment is incentive compatible. Since $p_{12}^* > 0$ always holds, the condition $\gamma_{12} > \tilde{\gamma}$ is necessary and sufficient to insure the existence of punishment in equilibrium. This concludes the proof of Theorem 1

A.1.1 Proof of Theorem 3

Without punishment, in period t, the equilibrium is such that $a_{i,t}^* = \tilde{a}_{i,t}(\mathbf{r_t})$, with $\tilde{a}_{i,t}(\mathbf{r_t})$ the contribution effort that maximizes (2), for any $i \in \mathcal{N}$.

At the end of the first period $t = 1$, individual i revises his beliefs on the social roles to maximize his utility:

$$\mathbf{r_{i,2}} = \arg\max_{\mathbf{r}} u_i(\tilde{a}_{i,1}(\mathbf{r}), \tilde{\mathbf{a}}_{-\mathbf{i,t}}(\mathbf{r})). \tag{A.11}$$

Hence, it is direct that

$$a_{s_i,2}(s_j) = \tilde{a}_{s_i,1} \tag{A.12}$$

for any $i \in \mathcal{N}$. That is, the optimal social beliefs on the social roles in period 2 correspond to the equilibrium behaviors that have been adopted by the agents in period 1.

Hence, the first-order condition associated with the determination of $a^*_{i,2}$ is:

$$\theta_i - a_i - \alpha \mid a_1 - \tilde{a}_{1,1} \mid = 0, \tag{A.13}$$

with $\alpha = n_1\alpha_{11} + n_2\alpha_{12}$, from which I deduce that

$$a^*_{i,2} = \frac{\theta_i + \alpha a^*_{i,1}}{1+\alpha}. \tag{A.14}$$

The same reasoning applies in period $t > 1$ and I find that

$$a^*_{i,t+1} = \frac{\theta_i + \alpha a^*_{i,t}}{1+\alpha}. \tag{A.15}$$

In the long-run, $a^*_{i,\infty}$ solves the fixed point equation

$$a^*_{i,\infty} = \frac{\theta_i + \alpha a^*_{i,\infty}}{1+\alpha}, \tag{A.16}$$

from which I deduce that

$$a^*_{i,\infty} = \theta_i. \tag{A.17}$$

Hence, given that

$$a_{i,t+1}(j) = \tilde{a}_{i,t} \tag{A.18}$$

for any $i, j \in \mathcal{N}$ from the maximization (3), we deduce that $a_{i,\infty}(j) = \theta_i$ for any $i, j \in \mathcal{N}$. This concludes the proof of Theorem 3.

A.1.2 Proof of Theorem 4

I define $\tilde{\gamma}_1$ as the threshold value of γ above which the individuals of type 1 punish off the equilibrium the individuals of type 2 in period 1 when they deviate from their assigned social role. Hence, given Theorem 1, since $\tilde{a}_{2,1} > 0 = a_{2,0}(1)$ and $a_{2,0}(2) = \theta_2$, we know from (A.7) that

$$\tilde{\gamma}_1 = \frac{2n_2\theta_2 - 1/2\tilde{a}_{2,1} - n_1\alpha_{21}/2\tilde{a}_{2,1} + n_2\alpha_{22}/2(2\theta_2 - \tilde{a}_{2,1})}{n_2/2\tilde{a}_{2,1}}, \tag{A.19}$$

which rewrites

$$\tilde{\gamma}_1 = \frac{2\theta_2(2+\alpha_{22})}{\tilde{a}_{2,1}} - \frac{1}{n_2}(1 + n_2\alpha_{22} - n_1\alpha_{21}), \tag{A.20}$$

with $\tilde{a}_{2,1}$ the most preferred contribution of the individuals of type 2 in period 1. Maximizing the utility of the agents of type 2 in period 1, we find that

$$\tilde{a}_{2,1} = \frac{\theta_2(1 + n_2\alpha_{22})}{1 + n_1\alpha_{21} + n_2\alpha_{22}}. \tag{A.21}$$

Consider first the case where $\gamma > \tilde{\gamma}_1$. Since $\gamma > \tilde{\gamma}_1$, the individuals of type 2 are punished in period 1 when they deviate from their equilibrium behavior $a_{2,1}(1) = 0$. Hence, they choose action $a^*_{2,1} = 0$. As a result, from maximization (3), individuals revise their beliefs on the social roles of the individuals of type 2 and $a_{2,2}(i) = 0$ for any $i \in \mathcal{N}$.

We deduce from the maximization (2) that if the individuals of type 2 are punished in period 1, they choose a contribution that maximizes their utility (2) with $a_{2,2}(i) = 0$ for any $i \in \mathcal{N}$. We find that they would choose

$$\tilde{a}_{2,2} = \frac{\theta_2}{1 + n_1\alpha_{21} + n_2\alpha_{22}}. \tag{A.22}$$

Hence, the threshold above which individuals of type 1 punish those of type 2 also changes in period 2. We denote $\tilde{\gamma}_2$ the threshold value of γ above which the individuals of type 1 punish those of type 2 in period 1. We find from (A.7) that

$$\tilde{\gamma}_2 = \frac{4\theta_2}{\tilde{a}_{2,2}} - \frac{1}{n_2}(1 + n_2\alpha_{22} - n_1\alpha_{21}). \tag{A.23}$$

Substituting (A.21) in the expression of $\tilde{\gamma}_1$ and (A.23) in the expression of $\tilde{\gamma}_2$,, we find that the inequality

$$\tilde{\gamma}_1 < \tilde{\gamma}_2 \tag{A.24}$$

rewrites

$$1 < 2n_2, \tag{A.25}$$

which is true. Hence the inequality $\tilde{\gamma}_1 < \tilde{\gamma}_2$ is necessarily verified. Hence, to complete the proof, we need to consider two subcases.

Case $\gamma > \tilde{\gamma}_2 > \tilde{\gamma}_1$. The individuals of type 2 remains punished in period 2 if they deviate from their assigned behavior. Hence, from period 3 on, the game remains identical to the game of period 2. We deduce the following result:

If $a^*_{i,t}(\mathbf{r}_t) = \tilde{a}_{i,t}(\mathbf{r_t})$ *if* $s_i = 1$ *and* $a^*_{i,t}(\mathbf{r}_t) = 0$ *if* $s_i = 2$ *and* $p^*_{12,t}(\mathbf{r_t}) > 0$ *in any period* t. *In the long-run,* $a_{i,\infty}(j) = a^*_{i,\infty} = \theta_j$ *if* $s_j = 1$, *and* $a_{i,\infty}(j) = a^*_{i,\infty} = 0$ *otherwise for any* $i \in \mathcal{N}$.

Case $\tilde{\gamma}_2 > \gamma > \tilde{\gamma}_1$. The contribution level of the individuals of type 2 is low enough in period 2 and they are not punished by the individuals of type 1. Hence, at the end of period 2, the agents revise their beliefs on the social roles and $a_{2,3}(i) = \tilde{a}_{2,2}$ for any $i \in \mathcal{N}$.

In period 3, if the individuals of type 2 were not punished, they would choose a contribuition that maximizes their utility (2). We find that they would choose

$$\tilde{a}_{2,3} = \frac{\theta_2 + \alpha \tilde{a}_{2,2}}{1+\alpha}, \tag{A.26}$$

with $\alpha = n_1\alpha_{21} + n_2\alpha_{22}$, from which it is direct that

$$\tilde{a}_{2,3} > \tilde{a}_{2,2}. \tag{A.27}$$

Hence, the threshold above which individuals of type 1 punish those of type 2 also changes in period 3. We denote $\tilde{\gamma}_3$ the threshold value of γ above which the individuals of type 1 punish those of type 2. We find from (A.7)

$$\tilde{\gamma}_3 = \frac{4\theta_2 n_2 - \tilde{a}_{2,2}}{n_2(\tilde{a}_{2,3} - \tilde{a}_{2,2})} - \frac{1}{n_2}(1 + \alpha_{11}n_1 + \alpha_{21}n_2). \tag{A.28}$$

Hence, the inequality $\tilde{\gamma}_3 > \tilde{\gamma}_2$ rewrites

$$\frac{4\theta_2 n_2 - \tilde{a}_{2,2}}{n_2(\tilde{a}_{2,3} - \tilde{a}_{2,2})} > \frac{4\theta_2}{\tilde{a}_{2,2}}, \tag{A.29}$$

which rewrites

$$4n_2 > 1. \tag{A.30}$$

This this last inequality is true, the inequality $\tilde{\gamma}_3 > \tilde{\gamma}_2$ is necessarily verified. If the individuals of type 2 are not punished in period 2 because $\gamma < \tilde{\gamma}_2$, they will not be punished in period 3, because $\gamma < \tilde{\gamma}_2 < \tilde{\gamma}_3$.

Applying the same reasoning for any period $t > 3$, we find that

$$\tilde{\gamma}_t = \frac{4\theta_2 n_2 - \tilde{a}_{2,t-1}}{n_2(\tilde{a}_{2,t} - \tilde{a}_{2,t-1})} - \frac{1}{n_2}(1 + \alpha_{11}n_1 + \alpha_{21}n_2). \tag{A.31}$$

so

$$\tilde{\gamma}_{t+1} > \tilde{\gamma}_t \tag{A.32}$$

is verified if

$$\frac{4\theta_2 n_2 - \tilde{a}_{2,t}}{n_2(\tilde{a}_{2,t+1} - \tilde{a}_{2,t})} > \frac{4\theta_2 n_2 - \tilde{a}_{2,t-1}}{n_2(\tilde{a}_{2,t} - \tilde{a}_{2,t-1})} \tag{A.33}$$

Given that

$$\tilde{a}_{2,t} = \frac{\theta_2 + \alpha \tilde{a}_{2,t-1}}{1+\alpha}, \tag{A.34}$$

we find that

$$\frac{4\theta_2 n_2 - \tilde{a}_{2,t}}{n_2(\tilde{a}_{2,t+1} - \tilde{a}_{2,t})} = \frac{4\theta_2 n_2 - \frac{\theta_2 + \alpha \tilde{a}_{2,t-1}}{1+\alpha}}{n_2 \frac{\theta_2 + \alpha \tilde{a}_{2,t} - \theta_2 - \tilde{a}_{2,t-1}}{1+\alpha}}, \tag{A.35}$$

so

$$\frac{4\theta_2 n_2 - \tilde{a}_{2,t}}{n_2(\tilde{a}_{2,t+1} - \tilde{a}_{2,t})} = \frac{4\theta_2 n_2 - \tilde{a}_{2,t-1}}{n_2(\tilde{a}_{2,t} - \tilde{a}_{2,t-1})} + \frac{\theta_2(4n_2 - 1)}{\alpha n_2(\tilde{a}_{2,t} - \tilde{a}_{2,t-1}} \tag{A.36}$$

from which it is direct that

$$\frac{4\theta_2 n_2 - \tilde{a}_{2,t}}{n_2(\tilde{a}_{2,t+1} - \tilde{a}_{2,t})} > \frac{4\theta_2 n_2 - \tilde{a}_{2,t-1}}{n_2(\tilde{a}_{2,t} - \tilde{a}_{2,t-1})}, \tag{A.37}$$

so

$$\tilde{\gamma}_{t+1} > \tilde{\gamma}_t \tag{A.38}$$

is verified for any $t > 3$. Given that $\tilde{a}_t$ increase over time after period 2, if the individual is not punished in period 2, he will not be punished in subsequent periods, as

$$\gamma < \tilde{\gamma}_2 < \tilde{\gamma}_3 < \cdots < \tilde{\gamma}_\tau \tag{A.39}$$

for any period $\tau > 3$. We deduce the following result:

$a^*_{i,0}(\mathbf{r}_t) = \tilde{a}_{i,t}(\mathbf{r_t})$ if $s_i = 1$ and $a^*_{i,0}(\mathbf{r}_t) = 0$ otherwise and $p^*_{12,0}(\mathbf{r_t}) > 0$. For any period $t > 1$, $a^*_{i,t}(\mathbf{r}_t) = \tilde{a}_{i,t}(\mathbf{r_t})$ for any $i \in \mathcal{N}$ and $p^*_{12,t}(\mathbf{r_t}) = 0$. Social roles and actions equalize abilities in the long-run, $a_{i,\infty}(j) = a^*_{i,\infty} = \theta_j$ for any $i, j \in \mathcal{N}$.

Finally, we consider the case where $\gamma < \tilde{\gamma}_1$. In this case, the individuals of type 1 are not punished initially, and they choose their most preferred contribution effort

$$\tilde{a}_{2,1} = \frac{\theta_2(1 + n_2\alpha_{22})}{1+\alpha}. \tag{A.40}$$

Hence, at the end of period 1, the agents revise their beliefs on the social roles and $a_{2,2}(i) = \tilde{a}_{2,1}$ for any $i \in \mathcal{N}$.

In period 2, if the individuals of type 2 were not punished, they would choose a contribution effort that maximizes their utility (2). We find that

$$\tilde{a}_{2,2} = \frac{\theta_2 + \alpha \tilde{a}_{2,1}}{1+\alpha}, \tag{A.41}$$

from which it is direct that

$$\tilde{a}_{2,2} > \tilde{a}_{2,1}. \tag{A.42}$$

The threshold above which individuals of type 1 punish those of type 2 also changes in period 2. We denote it $\tilde{\gamma}_2^N$. Following the same steps as above, I find that

$$\tilde{\gamma}_1 < \tilde{\gamma}_2^N \tag{A.43}$$

Hence, the individuals of type 2 will not be punished in period 2 if they have not been punished in period 1. The same reasoning can finally be applied for any period $t > 2$. We find that the threshold above which individuals of type 1 punish those of type 2 in period t, conditionally on not having punished before, $\tilde{\gamma}_t^N$ increases overtime. If the individuals of type 2 are not punished in period 1, they will not be punished in subsequent periods, as

$$\gamma < \tilde{\gamma}_1 < \tilde{\gamma}_2^N < \cdots < \tilde{\gamma}_\tau^N. \tag{A.44}$$

We deduce the following result: $a_{i,t}^*(\mathbf{r}_t) = \tilde{a}_{i,t}(\mathbf{r_t})$ *for any* $i \in \mathcal{N}$ *and* $p_{12,t}^*(\mathbf{r_t}) = 0$ *in any period* t. *Social roles and actions equalize abilities in the long-run,* $a_{i,\infty}(j) = a_{i,\infty}^* = \theta_j$ *for any* $i, j \in \mathcal{N}$.

This concludes the proof of Theorem 4.

A.1.3 Proof of Theorem 6

From (A.23), the threshold $\tilde{\gamma}_2$ can be rewritten

$$\tilde{\gamma}_2 = \frac{4\theta_2}{\tilde{a}_{2,2}} - \frac{1}{n_2}(1 + n_2\alpha_{22} - n_1\alpha_{21}). \tag{A.45}$$

with

$$\tilde{a}_{2,2} = \frac{\theta_2}{1 + n_1\alpha_{12} + n_2\alpha_{22}}. \tag{A.46}$$

Since $\tilde{a}_{2,2}$ decreases with n_2, $\tilde{\gamma}_2$ increases with n_2. We deduce that there exists a threshold $\tilde{n}_2$ such that if $n_2 < \tilde{n}_2$, $\gamma > \tilde{\gamma}_2$ and $\gamma \geq \tilde{\gamma}_2$ otherwise.

Hence, from (4), if $n_2 < \tilde{n}_2$, $W_\infty = (n - n_2)(n - \frac{1}{2})\theta^2$ and if $n_2 \geq \tilde{n}_2$, $W_\infty = n(n - \frac{1}{2})\theta^2$. This concludes the proof of Theorem 6.

A.1.4 Proof of Theorem 7

Let denote $\tilde{\gamma}_2(q)$ the threshold value of γ such that the individuals of type 1 punish those of type 2 in period 2 if $\gamma > \tilde{\gamma}_2(q)$ conditional on having punished them already in period 1 (see the Definition of the parameter in the proof of Theorem 4).

From the proof of Theorem 4, it is easy to show that the threshold $\tilde{\gamma}_2(q)$ is increasing in q. The higher the cost of punishing, the lower the set of parameter values such that punishment is optimal. Hence, there exists some threshold $\tilde{q} > 0$ such that if $q > \tilde{q}$, then $\gamma < \tilde{\gamma}_2(q)$ and there is no punishment in period 2.

If $q > \tilde{q}$, the beliefs on social roles change at the end of period 2 so as to reflect that the individuals of type 1 can produce. In period 3, from the proof of Theorem 4, $\tilde{\gamma}_2(q) < \tilde{\gamma}_3(q)$. Hence, by imposing again in period 3 a cost q on the punishers, the individuals of type 2 are still protected from punishment, as $\gamma < \tilde{\gamma}_2(q_2) < \tilde{\gamma}_3(q_2)$ and the beliefs on the social roles change accordingly.

The same reasoning holds for any period $t > 2$. Hence, there exists a period $\tau > 0$ such that

$$\tilde{\gamma}_{\tau-1}(0) < \gamma < \tilde{\gamma}_\tau(0). \tag{A.47}$$

From period τ on, since the punishment q has changed the beliefs on the social role of the individuals of type 2, it is not necessary anymore to impose a cost on the individuals of type 1 to force them not to punish the individuals of type 2. The beliefs on the social roles are such that punishment is not anymore necessary.

A.1.5 Proof of Theorem 8

In this extended version of the model,

$$p^*_{12,2} = u_2(a^*_{1,2}, \tilde{a}_{2,2}) - \epsilon - u_2(a^*_{1,2}, a_{2,2}(1)), \tag{A.48}$$

as the individuals of type 2 must be made indifferent between producing at their optimum and whistle-blowing and producing at the level $a_{2,2}(1)$. The incentive-compatibility

constraint accounts for the legal cost imposed on punishers and

$$p^*_{12,2} + q < u_1(a^*_{1,2}, \tilde{a}_{2,2}) - u_1(a^*_{1,2}, a_{2,2}(1)). \tag{A.49}$$

Let $\tilde{\gamma}_2(q, \epsilon)$ denote the threshold value of γ above which the individuals of type 1 punish those of type 2. Developing the previous equations, we then find that $\tilde{\gamma}_2(q, \epsilon)$ is increasing in q and decreasing in ϵ. Hence, there exists a threshold $\tilde{q}(\epsilon)$ with q increasing with ϵ such that $\gamma < \tilde{\gamma}_2(q, \epsilon)$.

If $q > \tilde{q}(\epsilon)$ in period 2, the beliefs on the social roles change at the end of that period to reflect that the individuals of type 2 can produce.

Applying the reasoning in the proof of Theorem 7, there exists a period $\tau(\epsilon) > 0$ such that

$$\tilde{\gamma}_{\tau(\epsilon)-1}(q(\epsilon) = 0) < \gamma < \tilde{\gamma}_{\tau(\epsilon)}(q(\epsilon) = 0). \tag{A.50}$$

Finally, I demonstrate that $\tau(\epsilon = 0) < \tau(\epsilon)$. By definition, in period τ,

$$\tilde{\gamma}_{\tau-1}(0) < \gamma < \tilde{\gamma}_{\tau}(0). \tag{A.51}$$

Since $\tilde{\gamma}_{\tau(\epsilon)-1}(0) < \tilde{\gamma}_{\tau-1}(0)$ there are two possible situations.

Either

$$\tilde{\gamma}_{\tau(\epsilon)-1}(0) < \tilde{\gamma}_{\tau-1}(0) < \gamma < \tilde{\gamma}_{\tau(\epsilon)}(0) < \tilde{\gamma}_{\tau}(0) \tag{A.52}$$

so $\tau(\epsilon) = \tau$, or

$$\tilde{\gamma}_{\tau(\epsilon)-1}(0) < \tilde{\gamma}_{\tau(\epsilon)}(0) < \gamma < \tilde{\gamma}_{\tau}(0) \tag{A.53}$$

so $\tau(\epsilon) > \tau$. Hence, $\tau(\epsilon) \geq \tau$ necessarily holds.

A.1.6 Proof of Theorem 9

Static. Solving first the static game in period 1, we find that the buyer choose to buy the good when $V_b > q$. Since V_b is uniformly distributed on $[0, 1]$, the likelihood of trade occurring is

$$\pi(q_1) = \begin{cases} 1 - q_1 \text{ if } q_1 \in [0, 1] \\ 0 \text{ if } q_1 > 1 \text{ and} \\ 1 \text{ otherwise.} \end{cases} \tag{A.54}$$

We deduce that when the buyer is active, she sets a price $q_1^* = 1/2$. Hence, absent punishment, she enters the market and chooses $a_{s,1} = 1$ when $c < 1/4 - (\alpha bs - \alpha_{ss})/2$ and chooses $a_{s,1} = 0$ otherwise.

If there is punishment in equilibrium, the buyer must be indifferent between choosing to be active in the market and face punishment or staying inactive, so

$$p_{bs,1} = \max(0, 1/4 - c - (\alpha_{bs} - \alpha_{ss})/2). \tag{A.55}$$

The punishment is incentive compatible for the buyer if

$$p_{bs,1} < \frac{\gamma}{2} - \max(V_b - 1/2, 0). \tag{A.56}$$

Hence, we deduce that the unique equilibrium can be characterized as follows:

- if $\gamma > \tilde{\gamma}_1$, then $p_{bs,1}^* = \max(0, 1/4 - c - (\alpha_{bs} - \alpha_{ss})/2)$, $a_{s,1}^* = 0$ and $a_{b,1}^* = 0$.
- If $\gamma \leq \tilde{\gamma}_1$, $p_{bs,1}^* = 0$, $a_{b,1}^* = 1$ when $c < 1/4 - (\alpha bs - \alpha_{ss})/2$ and $a_{s,1}^* = 1$ $a_{s,1}^* = 0$ otherwise, $q_1^* = 1/2$ and $a_{b,1}^* = 1$ if $V_b > 1/2$ and $a_{b,1}^* = 0$ otherwise.

$$\tilde{\gamma}_1 = \max(0, 1/4 - c - (\alpha_{bs} - \alpha_{ss})/2) + \max(V_b - 1/2, 0). \tag{A.57}$$

Dynamics. There are two cases to consider.

First, if $\gamma < \tilde{\gamma}_1$, then the seller is active when $c < 1/4 - (\alpha_{bs} - \alpha_{ss})/2$. In this case, after the first play of the game, then both the buyer and the seller revise their beliefs on the social role and perceive that the seller should be active in subsequent plays. As a result, the buyer will not further punish the buyer. The buyer is active from period 2 on, as $c < 1/4 - (\alpha_{bs} - \alpha_{ss})/2 < 1/4 + (\alpha_{bs} + \alpha_{ss})/2$.

Second, if $\gamma \geq \tilde{\gamma}_1$, then the buyer is punished in period 1 and remains inactive in that period. As a result, social roles change to reflect the first period equilibrium and $a_{s,2}(b) = a_{s,2}(s) = 0$. Solving the equilibrium in that case, following the steps of the resolution in period 1,

$$p_{bs,2} = \max(0, 1/4 - c - (\alpha_{bs} + \alpha_{ss})/2) \tag{A.58}$$

and the punishment is incentive-compatible when

$$p_{bs,2} < \frac{\gamma}{2} - \max(V_b - 1/2, 0), \tag{A.59}$$

so the equilibrium in period 2 can be characterized as follows:

- if $\gamma > \tilde{\gamma}_2$, then $p_{bs,2} = \max(0, 1/4 - c - (\alpha_{bs} + \alpha_{ss})/2)$, $a_{s,2} = 0$ and $a_{b,2} = 0$.
- If $\gamma \leq \tilde{\gamma}_2$, $p_{bs,2} = 0$, $a_{b,2} = 1$ when $c < 1/4 - (\alpha bs + \alpha_{ss})/2$ and $a_{s,2} = 1$ $a_{s,2} = 0$ otherwise, $q_2 = 1/2$ and $a_{b,2} = 1$ if $V_b > 1/2$ and $a_{b,2} = 0$ otherwise, with

$$\tilde{\gamma}_2 = \max(0, 1/4 - c - (\alpha_{bs} + \alpha_{ss})/2) + \max(V_b - 1/2, 0) \tag{A.60}$$

Comparing (A.57) with (A.60), it is direct that $\tilde{\gamma}_2 \leq \tilde{\gamma}_1$. I now summarize the previous findings:

- If $\gamma < \tilde{\gamma}_2$, the buyer is not punished in the two first plays of the game. He will not be punished in subsequent plays.
- If $\gamma \in [\tilde{\gamma}_2, \tilde{\gamma}_1]$, the buyer is not punished initially and will not be punished either in period 2.
- If $\gamma > \tilde{\gamma}_1$, the buyer is punished in the two first plays of the game and will be punished as well in all subsequent plays.

This concludes the proof of Theorem 9.

A.2 Extension with Prospective Internal Design of Beliefs on the Social Roles

I then assume that before playing in period $t+1$, the agents update their beliefs given a blend of what they should have believed to best confront period t (their experienced utility) and what they should believe to best confront period $t+1$ (their decision utility). I denote $\lambda \in [0, 1]$ the degree of prospective versus retrospective thinking in the formation of beliefs. Hence, the agents choose their beliefs on the social roles to maximize the following objective:

$$\mathbf{r_{i,t+1}} = \arg\max_{\mathbf{r_i}} (1-\lambda) u_i(a^*_{i,t}(\mathbf{r}), \mathbf{a^*_{-i,t}}(\mathbf{r})) + \lambda u_i(a^*_{i,t+1}(\mathbf{r}), \mathbf{a^*_{-i,t+1}}(\mathbf{r})), \tag{A.61}$$

with $\mathbf{r} = [\mathbf{r_1}, \ldots, \mathbf{r_n}]$. In the case where $\lambda = 0$, the agents are fully retrospective and they choose their beliefs on the social roles exactly as outlined in the main text. By contrast,

when $\lambda = 1$, the agents are completely prospective in their thinking and choose to equip themselves with the best possible beliefs to confront period $t+1$.

Under (A.61), we establish the following result:

Theorem 10 *In any period t and for any $i, j \in \mathcal{N}$, $a_{j,t+1}(i)$ is the unique solution of the fixed-point equation $a_{j,t+1}(i) = (1-\lambda)a^*_{j,t} + \lambda a^*_{j,t+1}$.*

The equation $a_{j,t+1}(i) = (1-\lambda)a^*_{j,t} + \lambda a^*_{j,t+1}$ defines a fixed-point equation, given that $a^*_{j,t+1}$, the equilibrium contribution of agent j in period $t+1$ is a function of $a_{j,t+1}(i)$, and is uniquely determined (from Theorem 1). The solution is unique, given that $a^*_{j,t+1}$ is a linear function of $a_{j,t+1}(i)$, with $\frac{\partial a^*_{j,t+1}}{\partial a_{j,t+1}(i)} < 1$. This point is demonstrated in more details below.

Consider first the case where $\lambda = 1$. In this case, $a_{j,t+1}(i) = a^*_{j,t+1}$ for any $i, j \in \mathcal{N}$, so all agents' beliefs on the social roles perfectly match the agents' equilibrium actions. Then it is direct that there will be no punishment in equilibrium, so $p^*_{ij} = 0$ for any $i, j \in \mathcal{N}$ and the optimal action of any agent i is $a^*_i = \theta_i$.

Consider now the case where $0 \leq \lambda < 1$, where λ can be arbitrarily low.

I first replicate the result when there is no punishment (Theorem 3). Consider individual i, with $s_i = 1$ for $i \in \mathcal{N}$. The derivative of (2) with respect to $a_{1,t}$ writes:

$$\frac{\partial u_1}{\partial a_{1,t}} = -a_{1,t} + \theta_1 - n_1\alpha_{11} \mid a_{1,t} - a_{1,t}(1) \mid -n_2\alpha_{12} \mid a_{1,t} - a_{1,t}(2) \mid, \tag{A.62}$$

Hence, given Theorem 10, $a^*_{t,1}$ solves

$$0 = -a^*_{1,t} + \theta_1 - (1-\lambda)\alpha \mid a^*_{1,t} - a^*_{1,t-1} \mid, \tag{A.63}$$

with $\alpha = n_1\alpha_{11} + n_2\alpha_{12}$, which gives

$$a^*_{1,t} = \frac{\theta_1 + (1-\lambda)\alpha a^*_{1,t-1}}{1 + (1-\lambda)\alpha}. \tag{A.64}$$

Hence in the long-run, the equilibrium action of individual i solves

$$a^*_{1,\infty} = \frac{\theta_1 + (1-\lambda)\alpha a^*_{1,\infty}}{1 + (1-\lambda)\alpha}, \tag{A.65}$$

which directly gives

$$a^*_{1,\infty} = \theta_1. \tag{A.66}$$

Following the same reasoning, we easily find that $a^*_{2,\infty} = \theta_2$. Hence, Theorem 3 holds in the case where the internal design of social role is both retrospective and prospective, with λ that can be arbitrarily low.

I now replicate the main result (Theorem 4). Consider that the individuals of type 1 can now punish these of type 2. In period 1, the individuals of type 1 will choose to punish those of type 2 if $\gamma > \tilde{\gamma}_1$, with $\tilde{\gamma}_1$ given by (A.57), exactly as in the case where there is no prospective thinking.

Things change in period 2, because the individuals revise their internalized beliefs on the social role at the end of period 1. Hence, from Theorem A, if the agent of type 2 are punished in period 1, then every agent $i \in \mathcal{N}$ will perceive that their social role is

$$a_{2,2}(i) = \lambda a^*_{2,2}, \tag{A.67}$$

with $\tilde{a}_{2,2}$ their equilibrium contribution in period 2. As long as $\lambda > 0$, if everyone expects that the next equilibrium will be such that the agents of type 2 contribute, then the agents can change accordingly their beliefs on the social role of the agents of type 2.

Hence, in period 2, the first-best contribution level of the individual of type 2 solves

$$\theta_2 - a_2 - \alpha(1-\lambda)a_2 = 0 \tag{A.68}$$

so

$$\tilde{a}_{2,2} = \frac{\theta_2}{1 + (1-\lambda)\alpha}. \tag{A.69}$$

Hence, I define the threshold $\tilde{\gamma}_2(\lambda)$, by analogy with $\tilde{\gamma}_2$, as the threshold above which, in period 2, the individuals of type 2 will be punished by those of type 1, given that they already faced punishment in period 1.

Hence, when $\gamma > \tilde{\gamma}_2(\lambda)$, the agents of type 1 punish those of type 2 if (i) they do action $\tilde{a}_{2,2}$ and (ii) everyone's beliefs on the social roles of the agents of type 2 are $a_{2,2}(i) = \lambda \tilde{a}_{2,2}$.

Relying on the same reasoning as in the proof of Theorem 4 and given (A.7), I find that

$$\tilde{\gamma}_2(\lambda) = \frac{4n_2\theta_2 - 2\lambda\tilde{a}_{2,2}}{n_2(1-\lambda)\tilde{a}_{2,2}} - \frac{1}{n_2}(1 + n_1\alpha_{21} + n_2\alpha_{22}). \tag{A.70}$$

Substituting (A.69) in the last expression and differentiating with respect to λ, I find that

$$\tilde{\gamma}_2(\lambda)' > 0, \tag{A.71}$$

Hence, $\tilde{\gamma}_2(\lambda) > \tilde{\gamma}_2(0)$. Since $\tilde{\gamma}_2(0)$ corresponds to the threshold $\tilde{\gamma}_2$ in the case where the agents' thinkign is only retrospective, as outlined in the main text (see equation (A.60)) and that we have established in equation (A.24) that

$$\tilde{\gamma}_1 < \tilde{\gamma}_2, \tag{A.72}$$

we deduce that

$$\tilde{\gamma}_1 < \tilde{\gamma}_2(0) < \tilde{\gamma}_2(\lambda). \tag{A.73}$$

Following then the same steps as the proof of Theorem 4, I deduce the following generalization:

Theorem 11 *There exist two thresholds $\tilde{\gamma}_1 < \tilde{\gamma}_2(\lambda)$ with $\tilde{\gamma}_2(\lambda)$ increasing with λ such that*

- *If $\gamma < \tilde{\gamma}_1$, $a^*_{i,t}(\mathbf{r}_t) = \tilde{a}_{i,t}(\mathbf{r_t})$ for any $i \in \mathcal{N}$ and $p^*_{12,t}(\mathbf{r_t}) = 0$ in any period t. Social roles and actions equalize abilities in the long-run, $a_{i,\infty}(j) = a^*_{i,\infty} = \theta_j$ for any $i, j \in \mathcal{N}$.*

- *If $\gamma > \tilde{\gamma}_2(\lambda)$, $a^*_{i,t}(\mathbf{r}_t) = \tilde{a}_{i,t}(\mathbf{r_t})$ if $s_i = 1$ and $a^*_{i,t}(\mathbf{r}_t) = 0$ if $s_i = 2$ and $p^*_{12,t}(\mathbf{r_t}) > 0$ in any period t. In the long-run, $a_{i,\infty}(j) = a^*_{i,\infty} = \theta_j$ if $s_j = 1$, and $a_{i,\infty}(j) = a^*_{i,\infty} = 0$ otherwise for any $i \in \mathcal{N}$.*

- *If $\tilde{\gamma}_1 < \gamma < \tilde{\gamma}_2(\lambda)$, then $a^*_{i,0}(\mathbf{r}_t) = \tilde{a}_{i,t}(\mathbf{r_t})$ if $s_i = 1$ and $a^*_{i,0}(\mathbf{r}_t) = 0$ otherwise and $p^*_{12,0}(\mathbf{r_t}) > 0$. For any period $t > 1$, $a^*_{i,t}(\mathbf{r}_t) = \tilde{a}_{i,t}(\mathbf{r_t})$ for any $i \in \mathcal{N}$ and $p^*_{12,t}(\mathbf{r_t}) = 0$. Social roles and actions equalize abilities in the long-run, $a_{i,\infty}(j) = a^*_{i,\infty} = \theta_j$ for any $i, j \in \mathcal{N}$.*

The only difference with Theorem 4 is that the agents of type 1 will punish the agents of type 2 for a lower set of parameter values (as $\tilde{\gamma}_2(\lambda)$ increases with λ). As their beliefs on the social roles are in part prospective, the agents of type 1 internalize that the agents of type 2 could be contributors. Hence, the salience of the social role needs to be larger than before so that the agents of type 1 still do not want to let the agents of type 2 produce.