Social Roles

Avner Seror*
Aix Marseille School of Economics

May 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 highlight the role of retrospective thinking and slippery slope arguments in explaining the persistence of social roles and consider a set of policy interventions aimed at breaking social roles when they deteriorate economic outcomes.

JEL Classification Numbers: D7, D9, C73, J15, J16, J7, Z1
Keywords: Social Roles, Identity, Economic Development, Endogenous Preferences, Gender

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

*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.

(2009), Bordalo et al. (2016)). While these approaches are relevant in many instances, 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 sufficiently 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.

Although I start with a static model, the main results of this analysis come from the dynamic model. In each period, adult agents have offspring and play a public good game. The children inherit the type of their parent but can choose their beliefs on the social roles. The process of belief formation is retrospective, as the children adopt the best possible beliefs on social roles given how their parents played. Adults can similarly revise their beliefs during their lifespan. I find that the agents' optimal beliefs in any given period will reflect the optimal contributions made in the previous period. Hence, there is a fundamental interdependence between social roles and economic decisions. As

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

the agents make optimal contribution decisions, they form beliefs on the social roles in the next period, which feed back into their preferences and affect their contribution decisions.

I characterize the pure-strategy Subgame Perfect Equilibria of the game and find that a "slippery slope" argument is central to the logic behind the existence of these equilibria. According to the theory, the agents can credibly commit to punish those that deviate from their assigned social roles when not doing so will lead to a sequence of events where social roles gradually change in a direction that they dislike and these changes will not opposed in the future if they are not opposed now (the slippery slope).

I show that the joint dynamics of social roles and economic contributions display two types of stationary states. In the first stationary state, agents' contributions and social roles reflect their ability to perform the tasks at hand. This stationary state is reached when the importance of others' abiding by their social roles is low. In the second stationary state, some agents are punished initially and continue to be punished in any period if they deviate from their assigned social role. As a result, the stationary state is such that some agents will be perceived by everyone, including themselves, as having social roles that do not match their abilities. This stationary state 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 gender roles: 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 equilibrium where punishment persists in the long-run is utility-dominated by the equilibrium where there is no punishment, meaning that any agent has a lower long-run utility in the former equilibrium. 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 utility-dominant 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 and may have offered a window of opportunity, likely increasing female labor force participation and decreasing the social exclusion of Blacks in the United States.

After the welfare and policy analysis, the model is extended to market and self-nomination games, which also describe important cases of social roles in application. I find that the results of the paper are robust to these games. Hence, more applied phenomena can be looked through the lens of the model, including female labor force participation, caste-based division of labor, labor market discrimination, participation in strategic decisions (e.g. self-nomination for leadership positions) among other outstanding instances of social roles affecting economic outcomes.

Finally, while the main analysis considers that the formation of beliefs on social roles is only retrospective, I acknowledge that this might be a self-motivated choice made by the agents.[2] Accordingly, I extend the model in order to account for an endogenous degree of prospective thinking as the agents might also form beliefs on social roles independently from the history of the game. I find that there is a complementarity between the ways the agents think about social roles. This principle have multiple applications discussed in the paper, including households locked in traditional beliefs about gender roles and populist politicians generating social divides by fostering retrospective thinking.

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

[2]On self-motivated beliefs, see Bénabou and Tirole (2002, 2003, 2004, 2011*a*).

economic outcomes.[3] It is typically assumed that when individuals associate themselves with groups, their preferences are such that they compare their own behavior with the average behavior in these groups (Shayo (2009)). One of the most novel aspect of this paper is to highlight that the formation of beliefs on social roles is the outcome of a Nash equilibrium involving many agents strategically choosing their beliefs rather than being driven by isolated individual choices.[4] 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).[5] 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.

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 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]

[3] 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.

[4] The model thus links to how Acemoglu and Jackson (2017) have modeled the evolution of social norms and law-breaking behaviors. See also the recent analysis of culture as a set of attributes by Acemoglu and Robinson (2021).

[5] 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)).

[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.

[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.

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 present several extensions of the model, while Section 7 concludes. All the proofs are contained in the online 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 self-nomination games and market games (Section 6). The game has three stages. In the first stage, agent i commits to a punishment strategy $p_{ij,t}$ on agent $j \neq i$, which depends on agent j's contribution effort. In the second 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 is $\theta_i a_i$. Exerting an effort a_i requires a quadratic cost $a_i^2/2$. In the final stage, agent i can impose the punishment $p_{ij} \geq 0$ on agent $j \neq i$. The assumption that the agents can commit to a punishment strategy in the first stage of the game is made for convenience and is relaxed in dynamic settings in Section 3.

The monetary payoff of agent i can be written as

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

as it is equal to the overall contribution minus the cost of effort and the punishments.

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. $\mathcal{N}_s$ is the set of type s agents, $s \in \{1, 2\}$. 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.[8]

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:

$$u_i(a_i, \mathbf{a_{-i}}, \mathbf{r}) = x(a_i, \mathbf{a_{-i}}) + s_i(a_i, \mathbf{a_{-i}}, \mathbf{r}), \tag{2}$$

with agent i's social payoff

$$s_i(a_i, \mathbf{a_{-i}}, \mathbf{r}) = -\sum_{j \in \mathcal{N}} \frac{\alpha_{ij}}{2}(a_i - a_i(j))^2 - \sum_{j \in \mathcal{N}} \frac{\gamma_{ij}}{2}(a_j - a_j(i))^2, \tag{3}$$

where $\alpha_{ij}, \gamma_{ij} \in [0, 1]$. 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 $a_j(i)$.

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

Through the first term on the right-hand side of (3), the utility function accounts for the influence of both *self-image* and *social-image* on economic actions.[9] Through the second term on the right-hand side of (3), 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.[10] 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

Given the structure of the game, an equilibrium is defined using the standard notion of pure-strategy Subgame Perfect Nash Equilibrium. Such an equilibrium characterizes the optimal contribution efforts and punishments administered by the agents and will be denoted $\{\{a_i^*\}_{i\in\mathcal{N}}; \{p_{ij}^*)\}_{i,j\in\mathcal{N}}\}$.

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 allows to focus on the main moving parts without altering the reasoning of the model. Finally, I assume that the punishment is administered is bilateral interactions and is not common knowledge in the group. Hence, I abstract from issues of strategic free-riding in punishment that would naturally arise if punishment was common knowledge (Bramoullé and Kranton (2007)).

[9] 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.

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

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

- *If* $\gamma < \tilde{\gamma}$, $a_i^* = \tilde{a}_i$ *for any* $i \in \mathcal{N}$, *with*

$$\tilde{a}_i = \frac{\theta_i + \alpha_{s_i 1} a_{s_i}(1) + \alpha_{s_i 2} a_{s_i}(2)}{1 + \alpha_{s_i 1} + \alpha_{s_i 2}}$$

  the contribution effort that maximizes (2) and $p_{ij}^* = 0$ *for any* $j \in \mathcal{N}$.

- *If* $\gamma \geq \tilde{\gamma}$, $a_i^* = \tilde{a}_i$ *for any* $i \in \mathcal{N}_1$ *and* $a_j^* = a_2(1)$ *for any* $j \in \mathcal{N}_2$ *and*

$$\begin{cases} p_{ij}^* > 0 \text{ when } j \text{ deviates from } a_j^* = a_2(1), \text{ and} \\ p_{ij}^* = 0 \text{ otherwise.} \end{cases}$$

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 a type 1 agent perceives the importance γ of type 2 agents fulfilling their social role as great, he will punish them if they deviate from it. The punishment threat must be such that type 2 agents are indifferent between conforming to their assigned social role and choosing their most preferred contribution effort. As shown in the appendix, such a punishment threat is incentive-compatible for type 1 agents in the first-stage of the game when γ is sufficiently high.

Despite being incentive-compatible in the first stage of the game, the punishment threat is not incentive-compatible in the last stage of the game when it is supposed to be implemented. Hence, one important limitation the static approach is that the credibility of the commitment to a punishment strategy must be assumed due to reasons external to the model itself. In dynamic settings, I demonstrate that the credibility assumption can be relaxed without altering the results.

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)).[11]

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

[11]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.

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, 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 generalization of the static setup. I consider an overlapping generations model with an infinite number of periods. Each generation consists of N agents. An agent born in period $t-1$ becomes adult at the beginning of period t. Adulthood lasts for two periods. In their first adulthood period, the agents play a public good game against the other agents belonging to their generation. In their second adulthood period, the agents are old and inactive. The timing of the game can be summarized as follows. At the end of period $t-1$, (i) the old adults die and (ii) each agent born in that period chooses to internalize beliefs on the social roles of all the agents of his generation, including himself. We denote $\mathbf{r_{i,t-1}} = \{a_{i,t-1}(j)\}_{j\in\mathcal{N}}$ an agent i's beliefs on the social roles when i is born in period $t-1$ and adult in periods t and $t+1$. Agent i's beliefs $\mathbf{r_{i,t-1}}$ are sticky, meaning that they cannot be changed by the agents once adulthood is reached. In period t, the agents born in period $t-1$ play a public good game with the other adults belonging to their generation and have a single offspring. The old agents die while those born in that period choose their beliefs $\mathbf{r_{i,t}}$ for any $i \in \mathcal{N}$. Period $t+1$ starts, and so forth.

As in the static model, the good game has three stages. In the first stage of the game, any agent i chooses a punishment strategy $p_{ij,t}$ on agent $j \neq i$, which depends on agent j's contribution effort. I relax the assumption that in the first stage of the game, the agents can commit to a punishment strategy. In the second stage of the game, agent i chooses his contribution effort $a_{i,t} \geq 0$ to the public good. Agent i's ability $\theta_i \in [0,1]$ is still assumed fixed. In the third stage of the game, agent i implements his punishment strategy.

In period t, the agents born in $t-1$ have a utility

$$W_i = u_i(a_{i,t}, \mathbf{a_{-i,t}}, \mathbf{r_{t-1}} \mid h_t) + \beta s_i(a_{i,t+1}, \mathbf{a_{-i,t+1}}, \mathbf{r_t} \mid h_{t+1}), \tag{4}$$

where $\beta \in [0,1]$ corresponds to the agents' time preferences, u_i is given by (2), s_i by (3) and h_t is the history of the game in period t, $h_t = \{\{a_{i,\tau}(j)\}_{i,j\in\mathcal{N}}, \{a_{i,\tau}\}_{i\in\mathcal{N}}, \{p_{ij,\tau}\}_{i,j\in\mathcal{N}}, \}_{\tau<t}$. In their old age, the agents are inactive so they only care about future contribution efforts matching their beliefs on social roles. Through (4), I assume that any agent is able to judge the future according to her future beliefs on the social roles. Under this assumption, recently labeled *perfect mindset flexibility* by Bernheim et al. (2021), the agents are time consistent. The model can equally be solved when the agents are not perfectly mindset flexible.[12]

The equilibrium is still defined using the notion of pure-strategy Subgame Perfect Nash Equilibrium and will be denoted $\{\{a_{i,t}(j)^*\}_{i,j\in\mathcal{N}}, \{a^*_{i,t}\}_{i\in\mathcal{N}}, \{p^*_{ij,t}\}_{i,j\in\mathcal{N}}, \}_{t=1}^{\infty}$ as it characterizes, for any period, each agent's (i) beliefs on the social roles, (ii) contribution effort and (iii) punishment strategy. Given the specification (2), in the second stage of the public good game and in any period t, there is a mapping between the prevailing beliefs on social roles and the optimal contribution efforts made by the agents in that period. I denote $a^*_{i,t}(\mathbf{r_{t-1}})$ the optimal contribution effort decided in the second stage of the public good game as a function of the prevailing beliefs on social roles $\mathbf{r_{t-1}} = \{\mathbf{r_{i,t-1}}\}_{i\in\mathcal{N}}$.[13]

3.2 Social Roles

I assume that at the end of any period t, the agents choose to internalized beliefs on the social roles given the experienced utility of the adults playing the public good game in that period. Hence, at the end of period $t-1$, agents' beliefs on the social roles solve the following optimization program:

$$\mathbf{r^*_{i,t}} = \arg\max_{\mathbf{r_i}} u_i(a^*_{i,t-1}(\mathbf{r_i}, \mathbf{r^*_{-i,t}}), \mathbf{a_{-i,t-1}}^*(\mathbf{r_i}, \mathbf{r^*_{-i,t}}), (\mathbf{r_i}, \mathbf{r^*_{-i,t}}) \mid h_t), \tag{5}$$

with $\mathbf{r^*_{-i,t}} = \{\mathbf{r^*_{j,t}}\}_{j\in\mathcal{N}\setminus i}$.

[12] In that case, the results of the paper become stronger, given that when the agents are not mindset flexible they tend to value more conformity to their initial beliefs on social roles.

[13] To alleviate the notations, I do not condition $a^*_{i,t}(\mathbf{r_{t-1}})$ on the history of the game up to period t. With the complete notations, $a^*_{i,t}(\mathbf{r_{t-1}} \mid h_t)$ with $h_\tau = \{\{a_{i,t}(j)^*\}_{i,j\in\mathcal{N}}, \{a^*_{i,t}\}_{i\in\mathcal{N}}, \{p^*_{ij,t}\}_{i,j\in\mathcal{N}}\}_{t<\tau}$ for any $\tau > 0$.

In this model, while social identifies are automatically inherited from one generation to the next, how the social roles are perceived by the agents belonging to the two social identities evolve over time. I focus the exposition of the model on a case where the internal design of beliefs on social roles is retrospective and reflects a mechanism where each agent evaluates how much better off herself or her parent would have been had she perceived everyone's social roles differently, including her own. That is, in each period, an adult observe the equilibrium strategies and choose to adopt beliefs on social roles that would have been best to confront the game she just played.

In each period, it is as if the optimal vector of beliefs on the social role $\mathbf{r}_\mathbf{t}^*$ was the outcome of a Nash equilibrium played by the young agents when they choose to internalize beliefs on the social roles. That is, any agent strategically chooses to internalize beliefs on social roles depending on what she expects others to believe. Hence, equation (5) formalizes an important novel aspect of this model, which is that the formation of beliefs on the social roles is the result of an equilibrium process.

I acknowledge that the internal design of beliefs on social roles might also be prospective, as young agents could try to adopt beliefs on social roles that maximize their expected utility in the period where they become adult. Doing so requires by definition a certain degree of prospective thinking, as the agents need to be able to think about the future equilibrium strategies and how they can be changed by their outlooks on social roles. In Section 6.2, I extend the model and account for an endogenous degree of prospective thinking.

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 (5), at the end of each period t, each young agent's perception of social roles will reflect the equilibrium actions of all adult agents living in that period. Indeed, at the end of period t, an agent (he) internalizes the fact that had his parent (she) perceived the social roles as equal to the equilibrium actions, she would not have suffered a utility cost due to self-image concerns (when her action deviated from what she believed was her own social role), nor a utility cost due to others' deviation from the social role she assigned them.

3.3 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 pure-strategy Subgame Perfect Equilibrium where* $a^*_{i,t} = \tilde{a}_{i,t}$ *with* $\tilde{a}_{i,t}$ *the contribution effort that maximizes (4). In the long run, social roles and contribution efforts reflect 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 (5), young agents of both types adapt their perception of social roles to the equilibrium actions chosen by their parents. 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 demonstrated 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.

3.4 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, 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 characterizes the equilibrium contribution efforts and punishment strategies.

Theorem 4 *There exists a unique equilibrium such that*

- *If* $\gamma < \tilde{\gamma}_1$, $a^*_{i,t} = \tilde{a}_{i,t}$ *with* $\tilde{a}_{i,t}$ *the contribution effort that maximizes (4) and* $p^*_{ij,t} = 0$ *for any* $i, j \in \mathcal{N}$ *and any period* t. *In the long run, social roles and contribution efforts reflect abilities,* $a_{i,\infty}(j) = a^*_{i,\infty} = \theta_j$ *for any* $i, j \in \mathcal{N}$.

- *If* $\gamma \geq \tilde{\gamma}_1$, *for any agent* $i \in \mathcal{N}_1$ *and* $j \in \mathcal{N}_2$, $a^*_{i,t} = \tilde{a}_{i,t}$, $a^*_{j,t} = 0$, *and*

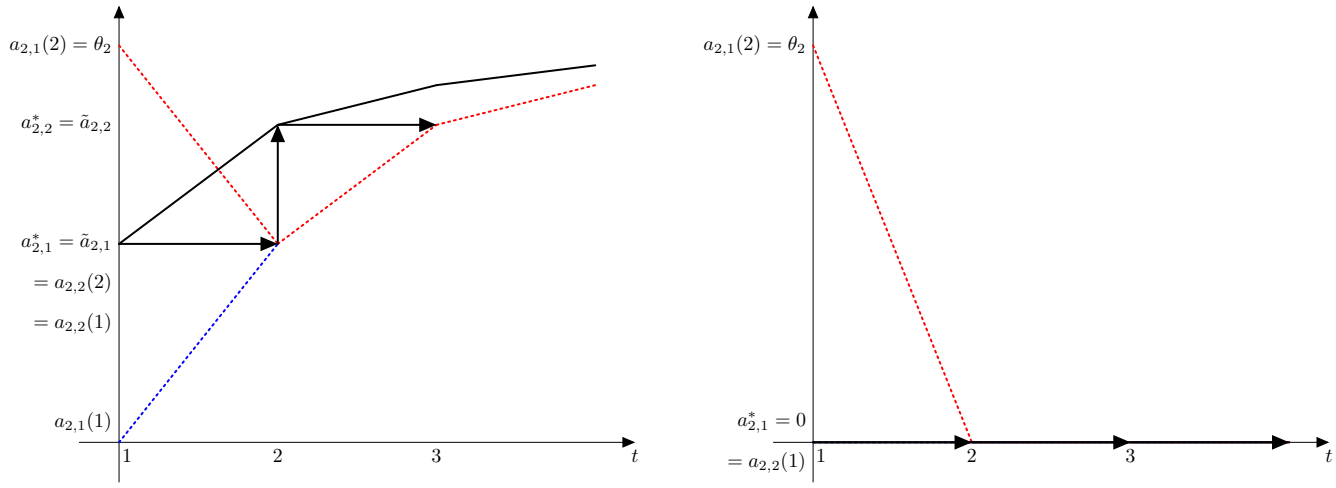
$$\begin{cases} p^*_{ij,t} > 0 \text{ when } j \text{ deviates from } a^*_{j,t} = 0, \text{ and} \\ p^*_{ij,t} = 0 \text{ otherwise.} \end{cases}$$

  in any period t. *In the long run,* $a_{i,\infty}(j)^* = a^*_{i,\infty} = \theta_i$ *and* $a_{j,\infty}(j)^* = a^*_{j,\infty} = 0$.

Combined, Theorems 2 and 4 characterize the unique pure-strategy SPE of the game. In any period t, type 1 agents can credibly impose a punishment threat on type 2 agents only if the latter remain punished in the next period. Otherwise, the punishment threat does not affect future social roles so it is not credible.[14] The main intuition behind Theorem 4 is then a "slippery slope" argument. According to Walton (2016), a key feature of the slippery slope argument is a progressive "loss of control" in a sequence of events in which each one event in the sequence causes the next one and so forth. According to the model, given that the type 1 agents are forward looking and adulthood lasts for more than one period of time, adults in their second period of life can credibly commit to punish those that deviate from their assigned social roles. Not doing so will lead to a sequence of events where social roles gradually change in a direction that they dislike and these changes will not opposed in the future if they are not opposed now (the slippery slope).

The slippery slope argument is commonly used to enforce social roles by justifying violence, discrimination or the opposition to progressive laws. For example, the abolition of slavery was commonly opposed in the Antebellum North on the ground that it would lead to a "miscegenation" of the American society. The slippery slope argument not only led to the enactment of laws supposed to preserve the so-called "racial purity", it also justified racial violence and discrimination against Blacks. More recently in France, a fierce opposition to same sex marriage legalization came from a conservative movement, "Manif Pour Tous" (or "Protest For All"). One of their key arguments was that same

[14] Theoretically, it could be that type 1 agents punish type 2 agents just to slow down the process of social roles changing. However, this does not materialize in equilibrium, as demonstrated in the proof of Theorem 4. Additionally, observe that Theorem 4 holds despite the agents being able to judge the future according to their future beliefs on the social roles. Hence, lack of mindset flexibility (Bernheim et al. (2021)) does not necessarily explain the persistence of social roles.



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

b) Case where $\gamma > \tilde{\gamma}_1$

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 $a^*_{t,2}$.

sex marriage legalization would drastically affect family values in the long run by opening the gate to medically assisted reproduction or surrogacy for same sex couples. Theorem 4 demonstrates that this logic is central to explain the persistence of social roles.

The joint evolution of social roles and economic behaviors outlined in Theorems 2 and 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 panel a), 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 b), only type 1 agents contribute to the common pool. Everyone believes that the 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 panel a) or the steady state of panels b) is reached hinges on the magnitude of γ. There are two 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. $\gamma > \tilde{\gamma}_1$). 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.

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.[15] 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

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

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.

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.[16] 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 equilib-

[16] 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.

rium, 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 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

In the rest of the paper, a long-run equilibrium will be characterized as more utility-dominant if all the agents have a higher utility in that equilibrium in period t with $t \to \infty$.[17]

Theorem 5 *The equilibrium reached when $\gamma > \tilde{\gamma}_1$ utility-dominates the equilibrium reached when $\gamma < \tilde{\gamma}_1$.*

When $\gamma > \tilde{\gamma}_1$, 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, in the equilibrium where

[17] While this utility-dominant concept is close to Pareto dominance, I prefer to eschew this concept as it is typically employed in a context where preferences are fixed.

$\gamma > \tilde{\gamma}_2$, 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 in the long-run, the equilibrium where $\gamma_2 > \tilde{\gamma}_1$ is utility-dominated, there is scope for public interventions that decrease the likelihood of reaching that 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 *Holding the number of agents in each generation N fixed, for any $\gamma > 0$, there exists a threshold number of type 2 agents $\tilde{n}_2 \in [0, N]$ and a threshold $\tilde{\tau} \in \mathbb{N}$ such that if $n_2 > \tilde{n}_2$ for at least $\tilde{\tau}$ periods of time, then the utility-dominated equilibrium is reached in the long run.*

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. If the quota is maintained for a sufficiently long period, beliefs regarding the social roles of type 2 agents change and the utility-dominated equilibrium outlined in Theorem 4 is not reached. This result strongly supports policies such as quotas and other forms of positive discrimination that temporarily 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 $\gamma > \tilde{\gamma}_1$, there exist a threshold $\tilde{q} > 0$ and a threshold $\tilde{\tau} \in \mathbb{N}$ such that if $q > \tilde{q}$ for at least $\tilde{\tau}$ periods of time, then the utility-dominated equilibrium is reached in the long run.*

When $\gamma > \tilde{\gamma}_1$, from Theorem 4, the equilibrium should be such that only type 1 agents contribute to the common pool in the long run. However, if society imposes for a finite period $\tilde{\tau}$ 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 $\tilde{\tau}$ 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 utility-dominated 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.

6 Extensions

The model can be extended in several directions.

6.1 Other 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. 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.[18]

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

[18] Formally, the following functional form for the monetary payoff $x(a_i, \mathbf{a_{-i}})$ can be written as $x(a_i, \mathbf{a_{-i}}) = \max_{j \in \mathcal{N}} \theta_j a_j$ 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 corresponds for example to experimental design of Coffman (2014).

should not self-nominate in these fields, as shown by Coffman (2014) in experimental settings.[19]

Market games. The model can also be extended to game settings where there is an exchange between two or more individuals. For simplicity, I consider in Appendix A.2 a version of the model where there are only two agents, a buyer and a seller. The overlapping structure of the game is similar to the main text. The only difference being that the adult agent play a market game. Indeed, in each period, 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. Both buyers and sellers form beliefs on their respective social roles. For example, it could be that buyers believe that it is not the social role of sellers to sell the good as sellers do not belong to the "right" caste, gender or race.[20]

Developing the dynamic model in a simple market game in Appendix A.2, I characterize the unique pure-strategy Subgame Perfect Equilibrium of the game. In particular, Appendix Theorem 9 is a generalization of Theorem 4 to a simple market game, as outlined above. This extension to market game describes how social roles 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.[21]

Finally, this model also 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

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

[20] In a seminal paper, Akerlof (1976) builds a static model describing a caste-based economy. Although the market game built in this extension is close to Akerlof (1976), I study the dynamics of social roles, an issue not discussed in the previous work. On the empirical side, Oh (2019) studies the effect of castes on economic decisions.

[21] 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.

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.[22]

6.2 The Origins of Retrospective Thinking

So far, I have assumed that when the young agents choose their beliefs on social roles, they only rely on retrospective thinking. Yet, the agents might be able to perceive future social roles independently from the past plays of the game. For certain purposes, it is therefore appropriate to extend to model to account for prospective thinking. Hence, I now assume that in the end of period t, the agents' beliefs on the social role solve the following optimization problem:

$$\mathbf{r_{i,t}} = \arg\max_{\mathbf{r_i}} (1-\lambda_{i,t}) u_i(a^*_{i,t-1}(\mathbf{r_i}, \mathbf{r^*_{-i,t}}), \mathbf{a_{-i,t-1}}^*(\mathbf{r_i}, \mathbf{r^*_{-i,t}}), (\mathbf{r_i}, \mathbf{r^*_{-i,t}})), + \\ \lambda_{i,t} u_i(a^*_{i,t}(\mathbf{r_i}, \mathbf{r^*_{-i,t}}), \mathbf{a_{-i,t-1}}^*(\mathbf{r_i}, \mathbf{r^*_{-i,t}}), (\mathbf{r_i}, \mathbf{r^*_{-i,t}})), \quad (6)$$

with $\mathbf{r^*_{-i,t}} = \{\mathbf{r^*_{j,t}}\}_{j \in \mathcal{N} \setminus i}$.

I interpret $\lambda_{i,t} \in [0,1]$ as a *degree of prospective thinking* of agent i in period t. When $\lambda_{i,t} = 0$, agent i is a retrospective thinker, in the sense that she only relies on the equilibrium strategies of the current generation to form her beliefs on the social roles. In contrast, when $\lambda_{i,t} = 1$, the agent is a prospective thinker, in the sense that she forms her beliefs on the social roles by envisioning the strategies that she and the other agents will play when adult. To give an example, young agents that belong to an oppressed group can either assimilate the oppression of their parents as their own condition and build a commensurate view of social roles in society (retrospective thinking), provided that they are expecting to be perceived as inferior by others too. Alternatively, they could also try to envision a future

[22] On business culture, see, for example, Hofstede (1994).

where their perception of their own social role is independent from the oppression suffered by their parents.

Many religions and political ideologies encourage a certain degree of retrospective thinking by emphasizing the "glory of the past", that the social "suppremacy" associated to a given identity should remain unaltered or that it is critical to traditional ideology that gender roles do not change (e.g. men should work and women should not). For certain purposes, it is therefore appropriate to treat $\lambda_{i,t}$ as endogenous. Here, we ask under which conditions the theory predicts that people should gravitate toward retrospective thinking rather than prospective thinking.

For the purpose of illustration, I will simplify by focusing on a case where there are only two agents playing a public good game. One type 1 agent and one type 2 agent. These agents are born in period 0, play a public good game in period 1 and the game ends. As in the static model, type 1 agents can commit in the first stage of the public good game to a punishment strategy. In period 0, the social roles are such that $a_{1,0}(1) = a_{1,0}(2) = a_{1,0}$ and $a_2(1) = a_2(2) = 0$. In words, everyone initially believes that type 2's social role is not to contribute and type 1's social role is to contribute an effort $a_{1,0}$. I further simplify by assuming that any agent $i \in \{1,2\}$ is either a prospective thinker and $\lambda_i = 1$ or a retrospective thinker and $\lambda_i = 0$. The agents first choose their degree of prospective thinking $\lambda_i \in \{0,1\}$ and then play a public good game as outlined in Section 2. A pure-strategy Subgame Perfect Equilibria of the game is defined as $\{\{\lambda_i^*\}_{i\in\{1,2\}}, \{a_i^*\}_{i\in\{1,2\}}, p_{12}^*\}$. We establish the following result:

Theorem 8

- *Provided that self-image concerns α_{ii}, $i \in \{1,2\}$, are sufficiently great, there exists a first equilibrium where $\lambda_1^* = \lambda_2^* = 1$, $a_i^* = \theta_i$ for any $i,j \in \{1,2\}$ and $p_{12}^* = 0$.*
- *Provided that social-image concerns α_{ij}, $i,j \in \{1,2\}$, $j \neq i$, are sufficiently great, if $\gamma \leq \tilde{\gamma}$ or if $\gamma > \tilde{\gamma}$ but $a_{10} > \theta_1$, there exists a second equilibrium where $\lambda_1^* = \lambda_2^* = 0$, $a_1^* = \frac{\theta_1 + (\alpha_{11}+\alpha_{12})a_{1,0}}{1+\alpha_{11}+\alpha_{12}}$, $a_2^* = 0$ and*

$$\begin{cases} p_{12}^* > 0 \text{ when the type 2 agent deviates from } a_2^* = 0, \text{ and} \\ p_{12}^* = 0 \text{ otherwise.} \end{cases}$$

Theorem 8 highlights the existence of a fundamental complementarity between the ways the two agents think about social roles. The complementarity is primarily explained

by the existence of self-image and social-image concerns (hence the conditions on the α_{ij}, $i, j \in \{1, 2\}$ in Theorem 8). If an agent cares about her social image, she is more inclined to follow the beliefs about her social roles of a retrospective thinker, as she wishes her action to be close to what is expected of her. Similarly, if she cares about her self-image, when the other agent is a prospective thinker, she will be more willing to be a prospective thinker too, as it will annihilate any cost due to self-image concerns. These intuitions underlie the existence of two equilibria: one where the agents are retrospective thinkers and one where they are prospective thinkers.

This result has several interesting applications, including the following:

(i) *Gender roles in a household.* One key example illustrating the logic behind Theorem 8 is that of gender roles in a household. Suppose that in the past, the social role of a type 1 agent (he) was to be a breadwinner and $a_{1,0} > \theta_1$ while the social role of a type 2 agent (she) was not to contribute and $a_{2,0} = 0$. In these conditions, both the man and the woman can be locked in retrospective thinking. When the man is a retrospective thinker, the woman does not contribute and is potentially subject to domestic violence if she does. She then becomes a retrospective thinker too because it enables her to lock the man into contributing as much as he can to the household (as $a_{1,0} > \theta_1$) and conform his beliefs on her social role. Reciprocally, given that the woman is a retrospective thinker, the man has no choice than being a retrospective thinker too in order to adopt a behavior that accords with the woman's beliefs on his social role.

(ii) *Populism.* Populist politicians often claim that they are the voice of a silent majority which social status has been deteriorated and that only them can restore. For example, in the 2016 US election, the Trump vote was correlated with areas dependent upon manufacturing sectors hit by the penetration of Chinese imports. The "leave votes" for the Brexit were also concentrated in areas characterized by low income or a historic dependence on manufacturing (Norris (2019)). According to the theory, a populist rhetoric can lock some agents into retrospective thinking and rationalize the use of various forms of punishment to enforce social roles (e.g., racial discrimination or xenophobia). Such populist rhetoric can then draw profound divide by making people in oppressed groups (e.g., racial minority or immigrants) embrace retrospective thinking too and internalize beliefs on social roles that breed more violence and economic inefficiencies.

(iii) *Immigrants' integration.* This model gives one explanation for potential failures of immigrants' integration in developed societies: retrospective thinking makes people prone to use various forms of violence to enforce social roles (e.g., discrimination against immigrants in order to decrease their access to job opportunities). Retrospective thinking from one segment of society, e.g., from immigrants or from natives, might lock the rest of society into retrospective thinking.

(iv) *Prospective thinking in political or religious ideologies.* Political or religious ideologies that advocate for societal changes can lead people to envision a future where social roles are different from their historical distribution. For instance, the "Rainbow nation" idea promoted by South African leaders after the end of the apartheid regime aimed at uniting South Africans by promoting shared cultural values that people belonging to different groups will be building together. Religious rituals too can affect prospective thinking. 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 and from the internalized beliefs that sustain it. According to the theory, people might respond to such political or religious ideologies by becoming prospective thinkers and changing their beliefs on social roles. Society at large might benefit from such evolution even if it is restricted to a segment of the population, given the complementarity in prospective thinking outlined in Theorem 8.

There are still unanswered questions on the origins of retrospective thinking. Although Theorem 8 highlights the existence of a complementarity in the ways the agents think, I do not address how the agents deal with strategic uncertainty. An extension of the model that refines the typology of equilibria might be particularly relevant in future research.[23]

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

[23] For a model that tackles the issue of strategic uncertainty in the context of cultural diversity, see, for example, Kets and Sandroni (2020). Bicchieri (2006) also offers a thorough account of coordination and culture.

roles and how social roles reciprocally feed back into internalized preferences and affect economic outcomes.

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, the utility of everyone is lower 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.

One key limitation of this model is that 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.[24] 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)).

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.

Acemoglu, Daron and James A. Robinson. 2021. Culture, Institutions and Social Equilibria: A Framework. NBER Working Papers 28832 National Bureau of Economic Research, Inc.

Acemoglu, Daron and Matthew O. Jackson. 2017. "Social Norms and the Enforcement of Laws." *Journal of the European Economic Association* 15(2):245–295.

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

[24] There is an emerging literature on identity choice. See, for example, Atkin, Colson-Sihra and Shayo (2021).

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. 1976. "The Economics of Caste and of the Rat Race and Other Woeful Tales." *The Quarterly Journal of Economics* 90(4):599–617.

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

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.

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.

Bicchieri, Cristina. 2006. *The grammar of society : the nature and dynamics of social norms*. New York, NY: Cambridge University Press.

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.

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.

Kets, Willemien and Alvaro Sandroni. 2020. "A Theory of Strategic Uncertainty and Cultural Diversity." *The Review of Economic Studies* 88(1):287–333.

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.

Norris, Pippa. 2019. *Cultural backlash : Trump, Brexit, and the rise of authoritarian populism*. Cambridge, United Kingdom New York, NY, USA: Cambridge University Press.

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.

Oh, Suanna. 2019. Does Identity Affect Labor Supply?

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

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.

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. 2021. Legalized Same-Sex Marriage and Coming Out in America: Evidence from Catholic Seminaries. AMSE Working Papers 2124 Aix-Marseille School of Economics, France.

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.

Walton, Douglas. 2016. *Slippery Slope*. Cham: Springer International Publishing pp. 2623–2632.

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:

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

so $\frac{\partial u_1}{\partial a_1}$ 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 $\tilde{a}_1$, with

$$\tilde{a}_1 = \frac{\theta_1 + \alpha_{11}a_1(1) + \alpha_{12}a_1(2)}{1 + \alpha_{11} + \alpha_{12}}. \tag{A.3}$$

A similar reasoning applies in the case of type 2 agents and we deduce that

$$\tilde{a}_2 = \frac{\theta_2 + \alpha_{21}a_2(1) + \alpha_{22}a_2(2)}{1 + \alpha_{21} + \alpha_{22}}. \tag{A.4}$$

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

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

Since we assume that type 1 agents can commit to a punishment strategy in the first stage of the public good game, the minimum punishment p_{12}^* is set by the individuals of type 1 so that type 2 agents are made indifferent between choosing their optimal action $\tilde{a}_2$ and conforming to the social role assigned to them. Hence, if there is a punishment in equilibrium, then

$$p_{12}^* = u_2(a_1^*, \tilde{a}_2) - u_2(a_1^*, a_2(1)). \tag{A.5}$$

Since type 1 agents can commit to a punishment strategy in stage 1, the punishment is incentive-compatible when

$$n_2 p_{12}^* < u_1(a_1^*, \tilde{a}_2) - u_1(a_1^*, a_2(1)). \tag{A.6}$$

We find that

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

and the incentive compatibility constraint for type 1 agents in the first stage of the game rewrites[25]

$$n_2 p_{12}^* < (\tilde{a}_2 - a_2(1))\{-n_2\theta_2 + n_2\gamma_{12}/2(\tilde{a}_2 - a_2(1))\}. \tag{A.8}$$

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

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

or

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

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

$$n_2\theta_2 - 1/2(\tilde{a}_2 + a_2(1)) - 1/2\alpha_{21}n_1(\tilde{a}_2 - a_2(1)) + n_2\alpha_{22}/2(2a_2(2) - a_2(1) - \tilde{a}_2) > -\theta_2 + \gamma_{12}/2(\tilde{a}_2 - a_2(1)) \tag{A.11}$$

or

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

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

[25] Notice that the punishment threat is not credible in the third stage of the game, when it is supposed to be implemented. This issue is inherent to the static settings of the model. We relax the assumption that type 1 agents can commit to a punishment strategy in Section **??**.

A.1.1 Proof of Theorem 3

Without punishment, in period t, the equilibrium is such that $a^*_{i,t} = \tilde{a}_{i,t}$, with $\tilde{a}_{i,t}$ 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:

$$\mathbf{r_{i,2}} = \arg\max_{\mathbf{r_i}} u_i(\tilde{a}_{i,1}(\mathbf{r_i}, \mathbf{r}^*_{-\mathbf{i,2}}), \tilde{\mathbf{a}}_{-\mathbf{i,t}}(\mathbf{r_i}, \mathbf{r}^*_{-\mathbf{i,2}}), (\mathbf{r_i}, \mathbf{r}^*_{-\mathbf{i,2}}). \tag{A.13}$$

Hence, it is direct that

$$a_{i,2}(j) = \tilde{a}_{i,1} \tag{A.14}$$

for any $i, j \in \mathcal{N}$. That is, the optimal 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:

$$\theta_i - a_i - \alpha_i(a_1 - \tilde{a}_{1,1}) = 0, \tag{A.15}$$

with $\alpha_i = n_1\alpha_{i1} + n_2\alpha_{i2}$, from which I deduce that

$$a^*_{i,2} = \frac{\theta_i + \alpha_i a^*_{i,1}}{1 + \alpha_i}. \tag{A.16}$$

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

$$a^*_{i,t+1} = \frac{\theta_i + \alpha_i a^*_{i,t}}{1 + \alpha_i}. \tag{A.17}$$

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

$$a^*_{i,\infty} = \frac{\theta_i + \alpha_i a^*_{i,\infty}}{1 + \alpha_i}, \tag{A.18}$$

from which I deduce that

$$a^*_{i,\infty} = \theta_i. \tag{A.19}$$

Hence, given that

$$a_{i,t+1}(j) = \tilde{a}_{i,t} \tag{A.20}$$

for any $i, j \in \mathcal{N}$ from the maximization (5), 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

Let $h_\tau = \{\{a_{i,t}(j)^*\}_{i,j\in\mathcal{N}}, \{a^*_{i,t}\}_{i\in\mathcal{N}}, \{p^*_{ij,t}\}_{i,j\in\mathcal{N}}, \}_{t<\tau}$ denote the history of the game in period τ. We denote $\mathbf{a_t^*} = \{a^*_{i,t}\}_{i\in\mathcal{N}}$, $\mathbf{b_t^*} = \{a_{i,t}(j)^*\}_{i,j\in\mathcal{N}}$, $\mathbf{p_t^*} = \{p^*_{ij,t}\}_{i,j\in\mathcal{N}}$ the equilibrium contribution, beliefs on social roles and punishment in period t.

From Theorem 3, since $\tilde{\mathbf{b}}_\mathbf{t} = \tilde{\mathbf{a}}_\mathbf{t}$ in any period $t \geq 1$, we can simplify the notation and denote $s_i(\mathbf{a_t^*}, \tilde{\mathbf{b}}^*_\mathbf{t} \mid h_t) = s_i(\mathbf{a_t^*} \mid h_t)$ and $u_i(\mathbf{a_t^*}, \tilde{\mathbf{b}}^*_\mathbf{t} \mid h_t) = u_i(\mathbf{a_t^*} \mid h_t)$.

Lemma 1 *There is no pure-strategy SPE such that $p^*_{ij,t} > 0$ but $p^*_{ij,t+1} = 0$.*

In order to demonstrate this result, we are going to proceed by strong induction.

Initialization. Consider the case where $t = 1$. Assume that in equilibrium, type 2 agents are not punished in period 2. A type 1 agent i faces the following IC constraint in period 1:

$$-n_2 p_{ij,1} + \beta u_1(\mathbf{a_2^*} \mid h_2) > \beta u_1(\mathbf{c_2^*} \mid h_2'), \tag{A.21}$$

with h_2 such that $a^*_{i,1} = \tilde{a}_{i,1}$ for any type 1 agent i, and $a^*_{i,1} = 0$ for any type 2 agent i. h_2' is such that $a^*_{i,1} = \tilde{a}_{i,1}$ for any $i \in \mathcal{N}$. By assumption, type 2 agents are not punished in period 2 so $\mathbf{a_2^*}$ is such that $a^*_{i,2} = \tilde{a}_{i,2}$ for any $i \in \mathcal{N}_1$ and $a^*_{j,2} = \tilde{a}_{j,1}$ for any $j \in \mathcal{N}_2$. Indeed, since type 2 agents are not punished in period 2 but were punished in period 1, they do their first-best effort $\tilde{a}_{j,1}$ in period 2. Similarly, c_2^* is such that $a^*_{i,2} = \tilde{a}_{i,2}$ for any $i \in \mathcal{N}$.

Substituting with (2), I find that the IC constraint can be rewritten as:

$$n_2 p_{ij,1} < \beta\{\theta_2(\tilde{a}_{2,1} - \tilde{a}_{2,2}) + \gamma/2[(\tilde{a}_{2,2} - \tilde{a}_{2,1})^2 - \tilde{a}^2_{2,1}]\}. \tag{A.22}$$

Intuitively, agent i understands that if he punishes agent j, he will postpone the process of social roles changing by one period. Hence, he needs to decides whether it is worth slowing down an inevitable change or simply let type 2 agents deviate from his beliefs on social roles.

Given that

$$\tilde{a}_{2,1} = \frac{\theta_2}{1+\alpha_2} \tag{A.23}$$

and

$$\tilde{a}_{2,2} = \frac{\theta_2 + \alpha_2 \tilde{a}_{2,1}}{1+\alpha_2} \tag{A.24}$$

with $\alpha_2 = n_1\alpha_{21} + n_2\alpha_{22}$, I find that inequality (A.22) cannot be verified.[26] Hence, it is not credible for a type 1 agent to punish a type 2 agent in period 1 if the equilibrium is such that there is no punishment threat on type 2 agents in period 2.

Strong induction. Assume that for any period $\tau < t$, it is true that if type 2 agents are not punished in period $\tau+1$, then they are not punished in period τ either. Under this assumption, we can demonstrate type 2 agents are not punished in period t, then they are not punished in period $t+1$.

Given the assumption of the strong induction reasoning, there is no history of the game depicting a SPE where there is a punishment in period $\tau < t$ but no punishment in period $\tau+1$. Hence, if type 2 agents have been punished in any period $\tau < t$, the history h_τ is necessarily such that $a^*_{2,\kappa} = 0$ for any $\kappa < \tau$, and the following IC constraint in period τ is verified:

$$n_2 p_{ij,\tau} < \beta\{-\theta_2\tilde{a}_{2,2} + \gamma(\tilde{a}_{2,2} - \tilde{a}_{2,1})^2\}. \tag{A.25}$$

Indeed, agent i trades-off punishing type 2 agents in period τ with breaking a sequence where $a^*_{2,\kappa} = 0$ for any $\kappa < \tau$ in period τ and letting type 2 agents do action $\tilde{a}_{2,1}$ in period τ and $\tilde{a}_{2,2}$ in $\tau+1$.

Hence, in period t, the IC constraint faced by a type 1 agent i is

$$n_2 p_{ij,t} < \beta\{\theta_2(\tilde{a}_{2,1} - \tilde{a}_{2,2} + \gamma[(\tilde{a}_{2,2} - \tilde{a}_{2,1})^2 - \tilde{a}^2_{2,1}]\}. \tag{A.26}$$

when he expects no punishment in period $t+1$. This inequality is necessarily false, as demonstrated before.

We have demonstrated that in period t, it cannot be that $p^*_{ij,t} > 0$ but $p^*_{ij,t} = 0$. This concludes the proof of the Lemma.

We have also demonstrated that in a pure-strategy SPE where there is a punishment threat implemented in period t by agent i, then the IC constraint necessarily writes:

$$n_2 p_{ij,t} < \beta\{-\theta_2\tilde{a}_{2,2} + \gamma(\tilde{a}_{2,2} - \tilde{a}_{2,1})^2\}. \tag{A.27}$$

Indeed, agent i trades-off punishing type 1 agents in period t with breaking a sequence where $a^*_{2,\tau} = 0$ for any $\tau < t$ and letting type 2 agents do action $\tilde{a}_{2,1}$ in period t and $\tilde{a}_{2,2}$ in

[26] One finds that (A.22) rewrites $p_{ij,1} < \beta\frac{\theta_2^2}{(1+\alpha_2)^2}\{-\alpha_2 - \frac{1+2\alpha_2}{(1+\alpha_2)^2}\gamma\}$, which cannot be verified for a positive punishment threat.

$t+1$. Since by assumption of the strong induction this inequality was verified up to period t, it remains verified in period t.

Consider a history h_t of the game where type 1 agents are punished in any period $\tau < t$. The optimal punishment in period t is set so that type 2 individuals are made indifferent between not contributing or contributing and getting punished.

$$p_{ij,t} = u_2(\mathbf{a}_\mathbf{t}^* \mid h_t) + \beta u_2(\mathbf{a}_{\mathbf{t+1}}^* \mid h_{t+1}) - u_2(\tilde{\mathbf{a}}_\mathbf{t} \mid h_t) - \beta u_2(\tilde{\mathbf{a}}_{\mathbf{t+1}} \mid h'_{t+1}), \tag{A.28}$$

with $h'_{t+1} = h_t \cup \{\tilde{\mathbf{a}}_\mathbf{t}, \mathbf{p_t} = \mathbf{0}, \tilde{\mathbf{b}}_\mathbf{t} = \tilde{\mathbf{a}}_\mathbf{t}\}$. Notice that we assumed above that if type 2 agents are not punished in period t, then they are not punished in period $t+1$ either. This result (Lemma 2) is demonstrated below.

Given that in history h_t, type 2 agents are punished in any period up to period t, we can rewrite the previous equality as

$$p_{ij,t} = \tilde{a}_{2,1}\{\theta_2 - 1/2\tilde{a}_{2,1}(1+\alpha_2)\} + \beta\{\theta_2\tilde{a}_{2,2} - \tilde{a}_{2,2}^2/2 - \alpha_2/2(\tilde{a}_{2,2} - \tilde{a}_{2,1})^2\}. \tag{A.29}$$

Substituting (A.22) in the previous equality, we deduce that if the pure-strategy SPE is such that there is a positive punishment threat in any period, then *for any period* t, the following inequality must be verified

$$\tilde{a}_{2,1}\{\theta_2 - 1/2\tilde{a}_{2,1}(1+\alpha_2)\} + \beta\{\theta_2\tilde{a}_{2,2} - \tilde{a}_{2,2}^2/2 - \alpha_2/2(\tilde{a}_{2,2} - \tilde{a}_{2,1})^2\} < \frac{1}{n_2}\beta\{-\theta_2\tilde{a}_{2,2} + \gamma/2(\tilde{a}_{2,2} - \tilde{a}_{2,1})^2\}, \tag{A.30}$$

which rewrite

$$\gamma > \tilde{\gamma}_1 \tag{A.31}$$

with $\tilde{\gamma}_1$ characterized by an equality between the RHS and the LHS of (A.30).

In a final step, we need to prove the following result:

Lemma 2 *There is no pure-strategy SPE such that* $p_{ij,t}^* = 0$ *but* $p_{ij,t+1}^* > 0$ *for any* $i \in \mathcal{N}_1$ *and* $j \in \mathcal{N}_2$ *and any* $t \geq 0$.

That is, there is no pure-strategy SPE such that the type 2 agents are not punished in period t but are punished in the subsequent period $t+1$ for any $t \geq 0$. To demonstrate this result, we again proceed by strong induction.

Initialization. In the case where $t = 1$, assume that $p_{ij,1}^* = 0$ but $p_{ij,2}^* > 0$. We are going to demonstrate that there is a contradiction. We are going to do this in three steps.

Step 1. Observe that given that $p^*_{ij,1} = 0$, $\gamma < \tilde{\gamma}_1$ is necessarily verified.

Step 2. We are going to demonstrate that $p^*_{ij,2} > 0$ is equivalent to the inequality

$$\gamma > \tilde{\gamma}_2. \tag{A.32}$$

Step 3. We are going to demonstrate that $\tilde{\gamma}_2 > \tilde{\gamma}_1$.

Since $p^*_{ij,1} = 0$ but $p^*_{ij,2} > 0$ is equivalent to $\gamma < \tilde{\gamma}_1 < \tilde{\gamma}_2 < \gamma$, there is a contradiction. Hence, we will have demonstrated that there is no pure-strategy SPE such that $p^*_{ij,1} = 0$ but $p^*_{ij,2} > 0$.

Here are the proofs of Steps 2 and 3.

The inequality $p^*_{ij,2} > 0$ is verified as long as

$$p^*_{ij,2} = u_2(\mathbf{a^*_2} \mid h_2) + \beta u_2(\mathbf{a^*_3} \mid h_3) - u_2(\tilde{\mathbf{a}}_\mathbf{2} \mid h_2) - \beta u_2(\tilde{\mathbf{a}}_\mathbf{3} \mid h'_3) \tag{A.33}$$

with

$$h_3 = h_2 \cup \{\{a^*_i = \tilde{a}_i, a^*_j = 0\}_{i\in\mathcal{N}_1, j\in\mathcal{N}_2}, \{p^*_{ij,2} > 0\}_{i\in\mathcal{N}_1, j\in\mathcal{N}_2}, \tilde{\mathbf{b}}_\mathbf{2} = \tilde{\mathbf{a}}_\mathbf{1}\}.$$

and

$$h'_3 = h_2 \cup \{\{a^*_i = \tilde{a}_i\}_{i\in\mathcal{N}}, \{p^*_{ij,2} = 0\}_{i,j\in\mathcal{N}}, \tilde{\mathbf{b}}_\mathbf{2} = \tilde{\mathbf{a}}_\mathbf{1}\}.$$

We can rewrite (A.33) as

$$p^*_{ij,2} = (\tilde{a}_{2,2} - \tilde{a}_{2,1})\{\theta_2 - 1/2(1+\alpha_2)(\tilde{a}_{2,1} + \tilde{a}_{2,2})\} + \beta\{(\tilde{a}_{2,3} - \tilde{a}_{2,1})(\theta_2 - 1/2(\tilde{a}_{2,3} + \tilde{a}_{2,1}\frac{\alpha_2}{2}(\tilde{a}_{2,3} - \tilde{a}_{2,1})^2\} \tag{A.34}$$

The punishment threat is credible in period 2 when it is incentive-compatible:

$$n_2 p^*_{ij,2} < \beta\{u_1(\mathbf{a^*_3} \mid h_3) - u_2(\tilde{\mathbf{a}}_\mathbf{3} \mid h'_3)\}, \tag{A.35}$$

which can be rewritten

$$n_2 p^*_{ij,2} < \beta\{-\theta_2(\tilde{a}_{2,2} - \tilde{a}_{2,1}) + \frac{\gamma}{2}(\tilde{a}_{2,3} - \tilde{a}_{2,2})^2\} \tag{A.36}$$

Substituting (A.34) in (A.36), we deduce that the resulting inequality can be written as

$$\gamma > \tilde{\gamma}(\tilde{a}_{2,1}). \tag{A.37}$$

Observe that $\tilde{\gamma}_2(0) = \tilde{\gamma}_1$. Furthermore, after a few lines of computations, we find that $\tilde{\gamma}_2$ increases with $\tilde{a}_{2,1}$, which implies that

$$\tilde{\gamma}(\tilde{a}_{2,1}) > \tilde{\gamma}_1 \tag{A.38}$$

given that $\tilde{a}_{2,1} = \frac{\theta_2}{1+\alpha_2} > 0$. This concludes the proof of Step 3.

We have demonstrated that there is no pure-strategy SPE such that $p^*_{ij,1} = 0$ but $p^*_{ij,2} > 0$ for any $i \in \mathcal{N}_1$ and $j \in \mathcal{N}_2$.

Strong induction. Assume that for any period $\tau < t-1$, there is no pure-strategy SPE with a history such that $p^*_{ij,\tau} = 0$ but $p^*_{ij,\tau+1} > 0$. We are going to demonstrate that there is no pure-strategy SPE such that $p^*_{ij,t-1} = 0$ but $p^*_{ij,t} > 0$.

Assume that there is a pure-strategy SPE, which history is such that $p^*_{ij,t-1} = 0$ but $p^*_{ij,t} > 0$. By assumption, the history of this SPE is such that $p^*_{ij,\tau} = 0$ for any $\tau < t-1$. Hence, in period t, the punishment is such that

$$p^*_{ij,t} = u_2(\mathbf{a^*_t} \mid h_t) + \beta u_2(\mathbf{a^*_{t+1}} \mid h_{t+1}) - u_2(\tilde{\mathbf{a}}_\mathbf{t} \mid h_{t+1}) - \beta u_2(\tilde{\mathbf{a}}_\mathbf{t+1} \mid h'_{t+1}) \tag{A.39}$$

with

$$h_{t+1} = h_t \cup \{\{a^*_{i,t} = \tilde{a}_{i,t}, a^*_{j,t} = 0\}_{i\in\mathcal{N}_1, j\in\mathcal{N}_2}, \{p^*_{ij,t} > 0\}_{i\in\mathcal{N}_1, j\in\mathcal{N}_2}, \tilde{\mathbf{b}}_\mathbf{t-1} = \tilde{\mathbf{a}}_\mathbf{t}\}.$$

and under the assumption of the strong induction, history h_t is such that type 2 agents have not been punished from period 1 to period $t-1$, meaning that

$$h_t = \{\{a^*_{i,t} = \tilde{a}_{i,\tau}\}_{i\in\mathcal{N}}, \{p^*_{ij,\tau} = 0\}_{i\in\mathcal{N}_1, j\in\mathcal{N}_2}, \tilde{\mathbf{b}}_{\tau-\mathbf{1}} = \tilde{\mathbf{a}}_\tau\}_{\tau\leq t-1}$$

and

$$h'_{t+1} = h_t \cup \{\{a^*_{i,t} = \tilde{a}_{i,t}\}_{i\in\mathcal{N}}, \{p^*_{ij,t} = 0\}_{i\in\mathcal{N}_1, j\in\mathcal{N}_2}, \tilde{\mathbf{b}}_\mathbf{t} = \tilde{\mathbf{a}}_\mathbf{t-1}\}.$$

Given histories h_t, h_{t+1} and h'_{t+1}, we can rewrite $p^*_{ij,t}$ as

$$p^*_{ij,t} = (\tilde{a}_{2,t} - \tilde{a}_{2,t-1})\{\theta_2 - 1/2(1+\alpha_2)(\tilde{a}_{2,t-1} + \tilde{a}_{2,t})\} + \beta\{(\tilde{a}_{2,t+1} - \tilde{a}_{2,t-1})(\theta_2 - 1/2(\tilde{a}_{2,t+1} + \tilde{a}_{2,t-1}\frac{\alpha_2}{2}(\tilde{a}_{2,t+1} - \tilde{a}_{2,t-1}) \tag{A.40}$$

The punishment threat is credible in period t when it is incentive-compatible:

$$n_2 p^*_{ij,t} < \beta\{u_1(\mathbf{a^*_{t+1}} \mid h_{t+1}) - u_2(\tilde{\mathbf{a}}_\mathbf{t+1} \mid h'_{t+1})\}, \tag{A.41}$$

which can be rewritten

$$n_2 p^*_{ij,t+1} < \beta\{-\theta_2(\tilde{a}_{2,t} - \tilde{a}_{2,t-1}) + \frac{\gamma}{2}(\tilde{a}_{2,t+1} - \tilde{a}_{2,t})^2\} \tag{A.42}$$

Substituting (A.40) in (A.42), we deduce that the resulting inequality can be written as

$$\gamma > \tilde{\gamma}(\tilde{a}_{2,t-1}). \tag{A.43}$$

After a few lines of computations, we find that $\tilde{\gamma}(\tilde{a}_{2,t-1})$ increases with $\tilde{a}_{2,t-1}$. As $\tilde{a}_{2,t-1}$ increases with t, we deduce that $\tilde{\gamma}(\tilde{a}_{2,t-1}) > \tilde{\gamma}_1$.

Hence, since $\gamma < \tilde{\gamma}_1$ and $\tilde{\gamma}(\tilde{a}_{2,t-1}) > \tilde{\gamma}_1$, then $\gamma < \tilde{\gamma}(\tilde{a}_{2,t-1})$, implying that there is no punishment in period t.

We have demonstrated that there is no pure-strategy SPE such that $p^*_{ij,t} = 0$ but $p^*_{ij,t+1} > 0$ for any $i \in \mathcal{N}_1$ and $j \in \mathcal{N}_2$ and any $t \geq 0$. This concludes the proof of the Lemma.

To summarize, we have demonstrated that in a pure-strategy SPE, there are only two possible histories:

- if $\gamma > \tilde{\gamma}_1$, a punishment threat $p^*_{ij,t}$ characterized in equation (A.29) is implemented in any period t. Moreover, type 1 agents do their first-best action $\tilde{a}_{1,t}$,

$$a^*_{1,t} = \tilde{a}_{1,t} = \frac{\theta_1 + \alpha_1 \tilde{a}_{1,t-1}}{1 + \alpha_1} \tag{A.44}$$

  with $\alpha_1 = n_1\alpha_{11} + n_2\alpha_{12}$ while type 2 agents do not contribute in any period t, $a^*_{2,t} = 0$. Social roles can directly be deduced from equilibrium contribution using Theorem 3, $a^*_{i,t+1}(j) = a^*_{i,t}$ for any $i, j \in \mathcal{N}$.

- if $\gamma < \tilde{\gamma}_1$, a punishment threat $p^*_{ij,t}$ is not credible in any period t. Hence, type 2 agents are never punished. Equilibrium actions are such that

$$a^*_{i,t} = \tilde{a}_{i,t} = \frac{\theta_i + \alpha_i \tilde{a}_{i,t-1}}{1 + \alpha_i} \tag{A.45}$$

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

This concludes the proof of Theorem 4.

A.1.3 Proof of Theorem 6

For simplification purposes, I make the following assumption: $\alpha_{21} = \alpha_{22} = \alpha_2$. Under these assumption, $\tilde{a}_{2,2}$ and $\tilde{a}_{2,1}$ are independent from n_2 so from (A.30), the threshold $\tilde{\gamma}_1$ is increasing with n_2.

We deduce that there exists a threshold $\tilde{n}_2 \in [0, N]$ such that if $n_2 < \tilde{n}_2$, $\gamma > \tilde{\gamma}_2$ and $\gamma \geq \tilde{\gamma}_2$ otherwise. Hence, the utility-dominated equilibrium is reached if and only if $n_2 < \tilde{n}_2$. This concludes the proof of Theorem 6.

Assume that a quota is maintained for t periods of time. Let $n_2 + q$ be the number of type 2 agents in the group, with $q \leq 0$. Hence,

$$(n_2 + q)p^*_{ij,\tau} > \beta\{u_1(\mathbf{a}^*_{\tau+\mathbf{1}} \mid h_{\tau+1}) - u_2(\tilde{\mathbf{a}}_{\tau+\mathbf{1}} \mid h'_{\tau+1})\}, \tag{A.46}$$

for any $\tau \in \{1, \ldots, t\}$. In any period $\tau \in \{1, \ldots, t\}$, punishment would have been implemented by type 1 agents absent the quota,

$$n_2 p^*_{ij,\tau} < \beta\{u_1(\mathbf{a}^*_{\tau+\mathbf{1}} \mid h_{\tau+1}) - u_2(\tilde{\mathbf{a}}_{\tau+\mathbf{1}} \mid h'_{\tau+1})\}, \tag{A.47}$$

with

$$n_2 p^*_{ij,\tau} = u_2(\mathbf{a}^*_\tau \mid h_\tau) + \beta u_2(\mathbf{a}^*_{\tau+\mathbf{1}} \mid h_{\tau+1}) - u_2(\tilde{\mathbf{a}}_\tau \mid h_{\tau+1}) - \beta u_2(\tilde{\mathbf{a}}_{\tau+\mathbf{1}} \mid h'_{\tau+1}) \tag{A.48}$$

and $h_{\tau+1}$ a history of the game such that there is no punishment in any period $\kappa < \tau + 1$ and agents do their first-best contribution effort.

Applying the same reasoning as in the proof of THeorem 4, I find that the inequality (A.47) is then equivalent to

$$\gamma < \tilde{\gamma}(\tilde{a}_{\tau-1}), \tag{A.49}$$

where $\tilde{\gamma}(\tilde{a}_{\tau-1})$ is increasing with $\tilde{a}_{\tau-1}$. As $\tilde{a}_{\tau-1}$ increases with τ, we deduce that there exists a threshold $\tilde{\tau}$ such that

$$\tilde{\gamma}(\tilde{a}_{\tilde{\tau}-1}) < \gamma < \tilde{\gamma}(\tilde{a}_{\tilde{\tau}}), \tag{A.50}$$

meaning that conditional on not being punished for $\tilde{\tau}$ periods, type 1 agents' beliefs on the social roles change sufficiently to make punishment not incentive compatible. By contrast if the quota is removed in any period $t < \tilde{\tau}$, then type 1 agents will punish type 2 agents in equilibrium, as $\tilde{\gamma}(\tilde{a}_{\tilde{\tau}-1}) < \gamma$. This concludes the proof of Theorem 7.

A.1.4 Proof of Theorem 7

It is clear that if punishing a type 2 agent implies a sufficiently high cost q, then punishment is not incentive compatible for type 1 agents anymore. We are not going to characterize the minimum punishment level $\tilde{q} > 0$ in this proof but simply characterize the number of periods during which the cost q must be maintained so that social roles change and punishment does not remain an optimal strategy for the type 1 agents.

Assume that the punishment technology q is maintained for t periods of time. Hence,

$$n_2 p^*_{ij,\tau} + q > \beta\{u_1(\mathbf{a}^*_{\tau+\mathbf{1}} \mid h_{\tau+1}) - u_2(\tilde{\mathbf{a}}_{\tau+\mathbf{1}} \mid h'_{\tau+1})\}, \tag{A.51}$$

for any $\tau \in \{1, \ldots, t\}$. In any period $\tau \in \{1, \ldots, t\}$, punishment would have been implemented if type 1 agents were not threaten by the technology q when

$$n_2 p^*_{ij,\tau} < \beta\{u_1(\mathbf{a}^*_{\tau+\mathbf{1}} \mid h_{\tau+1}) - u_2(\tilde{\mathbf{a}}_{\tau+\mathbf{1}} \mid h'_{\tau+1})\}, \tag{A.52}$$

with

$$n_2 p^*_{ij,\tau} = u_2(\mathbf{a}^*_\tau \mid h_\tau) + \beta u_2(\mathbf{a}^*_{\tau+\mathbf{1}} \mid h_{\tau+1}) - u_2(\tilde{\mathbf{a}}_\tau \mid h_{\tau+1}) - \beta u_2(\tilde{\mathbf{a}}_{\tau+\mathbf{1}} \mid h'_{\tau+1}) \tag{A.53}$$

and $h_{\tau+1}$ a history of the game such that there is no punishment in any period $\kappa < \tau + 1$ and agents do their first-best contribution effort.

Again, applying a reasoning similar to the proof of Theorem 4, I find that the inequality (A.52) is then equivalent to

$$\gamma < \tilde{\gamma}(\tilde{a}_{\tau-1}), \tag{A.54}$$

where $\tilde{\gamma}(\tilde{a}_{\tau-1})$ is increasing with $\tilde{a}_{\tau-1}$. As $\tilde{a}_{\tau-1}$ increases with τ, we deduce that there exists a threshold $\tilde{\tau}$ such that

$$\tilde{\gamma}(\tilde{a}_{\tilde{\tau}-1}) < \gamma < \tilde{\gamma}(\tilde{a}_{\tilde{\tau}}), \tag{A.55}$$

meaning that conditional on not being punished for $\tilde{\tau}$ periods, type 1 agents' beliefs on the social roles change sufficiently to make punishment not incentive compatible. By contrast if the technology q is removed in any period $t < \tilde{\tau}$, then type 1 agents will punish type 2 agents in equilibrium, as $\tilde{\gamma}(\tilde{a}_{\tilde{\tau}-1}) < \gamma$. This concludes the proof of Theorem 7.

A.1.5 Proof of Theorem 8

I denote $v_i(\lambda_1, \lambda_2)$ the indirect utility of agent i in the first stage of the game, given that both the optimal contributions and the punishment threats are substituted in the utility function u_i. In this proof, we are going to characterize under which conditions both $\lambda_1^* = \lambda_2^* = 1$ and $\lambda_1^* = \lambda_2^* = 0$ are Nash equilibria of the first stage of the game.

Equilibrium with Prospective Thinking. We are first going to prove that $(\lambda_1^* = 1, \lambda_2^* = 1)$ is an equilibrium outcome. For that purpose, given that $\lambda_i \in \{0, 1\}$, we simply need to demonstrate that $\lambda_i = 1$ is a best-response to $\lambda_j = 1$ for any $i, j \in \{1, 2\}$, $i \neq j$.

$$v_1(1, 1) = \theta_1^2/2 + \theta_2^2, \tag{A.56}$$

given that $a_1^* = \theta_1$ and $a_2^* = \theta_2$ when the two agents are prospective thinkers. When evaluating $v_1(0, 1)$, there are two cases to consider. In the first case, the type 2 agent is punished ($\gamma > \tilde{\gamma}$), with $\gamma > \tilde{\gamma}$ the threshold characterized in the proof of Theorem 1 given that $\lambda_1 = 0$. In the second, he is not when the type 1 agent is a retrospective thinker (i.e., $\gamma \leq \tilde{\gamma}$).

When $\gamma > \tilde{\gamma}$,

$$v_1(0, 1) = \theta_1 a_1^* - a_1^{*2}/2 + \theta_2.0 - \alpha_{11}/2(a_1^* - a_{1,0})^2 \tag{A.57}$$

given that $a_2^* = 0$ when $\gamma > \tilde{\gamma}$, with

$$a_1^* = \frac{\theta_1 + \alpha_{11} a_{1,0}}{1 + \alpha_{11}}. \tag{A.58}$$

Hence, it is clear that $v_1(1, 1) > v_1(0, 1)$, given that the monetary payoff's highest value is $\theta_1^2/2$.

When $\gamma \leq \tilde{\gamma}$,

$$v_1(0, 1) = \theta_1 a_1^* - a_1^{*2}/2 + \theta_2 a_2^* - \alpha_{11}/2(a_1^* - a_{1,0})^2 \tag{A.59}$$

given that $a_2^* = \frac{\theta_2}{1+\alpha_{21}} < \theta_2$. Again, we deduce that $v_1(1, 1) > v_1(0, 1)$. Hence, it is always a best-response to be a prospective thinker for a type 1 agent when the type 2 agent is a prospective thinker.

Similarly for the type 2 agent:

$$v_2(1,1) = \theta_1^2 + \theta_2^2/2. \tag{A.60}$$

and

$$v_2(1,0) = \theta_1 a_1^* + \theta_2 a_2^* - a_2^{*2}/2 - \alpha_{22}/2 a_2^{*2} \tag{A.61}$$

given that when $\lambda_1 = 1$, the type 1 agent never punishes the type 2 agent and

$$a_1^* = \frac{\theta_1 + \alpha_{12} a_{1,0}}{1 + \alpha_{12}} \text{ and } a_2^* = \frac{\theta_2}{1 + \alpha_{22}}. \tag{A.62}$$

I find that

$$v_2(1,1) - v_2(1,0) = \theta_1 \alpha_{12} \frac{\theta_1 - a_{1,0}}{1 + \alpha_{12}} + \frac{1}{2} \frac{\alpha_{22} \theta_2^2}{1 + \alpha_{22}} \tag{A.63}$$

so $v_2(1,1) - v_2(1,0) > 0$ is necessarily true when $\theta_1 \geq a_{1,0}$. This means that $(\lambda_1^* = 1, \lambda_2^* = 1)$ is necessarily an equilibrium outcome when $\theta_1 \geq a_{1,0}$.

When $\theta_1 < a_{1,0}$, $v_2(1,1) - v_2(1,0) > 0$ is also verified for any value of α_{21} as long as

$$\alpha_{22} > \alpha_{22}^* \tag{A.64}$$

where α_{22}^* is defined by the following equality:

$$-\theta_1(a_{10} - \theta_1) + \frac{\alpha_{22} \theta_2^*}{2(1 + \alpha_{22})} = 0. \tag{A.65}$$

Hence, we have demonstrated that when a_{10} is below the threshold $\frac{\theta_2}{2\theta_1} + \theta_1$ and $\alpha_{22} > \alpha_{22}^*$, then $(\lambda_1^* = 1, \lambda_2^* = 1)$ is necessarily an equilibrium outcome for any value of γ.

Equilibrium with Retrospective Thinking.

Consider first the case of the type 2 agent.

When $\gamma > \tilde{\gamma}$:

$$v_2(0,0) = \theta_1 \overline{a}_1^* \tag{A.66}$$

with

$$\overline{a}_1^* = \frac{\theta_1 + (\alpha_{12} + \alpha_{11}) a_{10}}{1 + \alpha_{11} + \alpha_{12}} \tag{A.67}$$

and

$$v_2(0,1) = \theta_1 \underline{a}_1^* \tag{A.68}$$

with

$$\underline{a}_1^* = \frac{\theta_1 + \alpha_{12} a_{10}}{1 + \alpha_{12}} \tag{A.69}$$

Hence,

$$v_2(0,0) - v_2(0,1) > 0 \tag{A.70}$$

if and only if

$$\overline{a}_1^* > \underline{a}_1^*, \tag{A.71}$$

which is true if and only if

$$a_{10} > \theta_1 \tag{A.72}$$

We have demonstrated that the condition $a_{10} > \theta_1$ is necessary and sufficient for retrospective thinking from a type 2 agent to be a best-response to a retrospective thinking type 1 agent when $\gamma > \tilde{\gamma}$.

When $\gamma \leq \tilde{\gamma}$:

$$v_2(0,0) = \theta_1 \underline{a}_1^* + \theta_2 \underline{a}_2^* - \underline{a}_2^{*2}/2 - 1/2(\alpha_{21} + \alpha_{22})\underline{a}_2^{*2} \tag{A.73}$$

with

$$\underline{a}_1^* = \frac{\theta_1 + (\alpha_{12} + \alpha_{11}) a_{10}}{1 + \alpha_{11} + \alpha_{12}} \text{ and } \underline{a}_2^* = \frac{\theta_2}{1 + \alpha_{21} + \alpha_{22}} \tag{A.74}$$

and

$$v_2(0,1) = \theta_1 \overline{a}_1^* + \theta_2 \overline{a}_2^* - \overline{a}_2^{*2}/2 - 1/2\alpha_{21}\overline{a}_2^{*2} \tag{A.75}$$

with

$$\overline{a}_1^* = \frac{\theta_1 + \alpha_{11} a_{10}}{1 + \alpha_{11}} \text{ and } \overline{a}_2^* = \frac{\theta_2}{1 + \alpha_{21}} \tag{A.76}$$

After a few lines of calculus, I find that the inequality

$$v_2(0,0) - v_2(0,1) > 0 \tag{A.77}$$

is true if and only if $\alpha_{21} > \overline{\alpha}_{21}$ with $\overline{\alpha}_{21} > 0$. Hence, when $\gamma \leq \tilde{\gamma}$, retrospective thinking is a best-response for a type 2 agent to retrospective if and only if $\alpha_{21} > \overline{\alpha}_{21}$.

Consider now the case of a type 1 agent.

When $\gamma > \tilde{\gamma}$,

$$v_1(0,0) = \theta_1 \overline{a}_1^* - \overline{a}_1^{*2}/2 - 1/2(\alpha_{11} + \alpha_{12})(\overline{a}_1^* - a_{1,0})^2 \tag{A.78}$$

with

$$\overline{a}_1^* = \frac{\theta_1 + (\alpha_{11} + \alpha_{12})a_{1,0}}{1 + \alpha_{11} + \alpha_{12}}. \tag{A.79}$$

and

$$v_1(1,0) = \theta_1 \underline{a}_1^* - \underline{a}_1^{*2}/2 - 1/2\alpha_{12}(\underline{a}_1^* - a_{1,0})^2 \tag{A.80}$$

with

$$\underline{a}_1^* = \frac{\theta_1 + \alpha_{11})a_{1,0}}{1 + \alpha_{11}}. \tag{A.81}$$

Hence, we can establish that

$$v_1(0,0) - v_1(1,0) > 0 \text{ if and only if } \alpha_{12} > \underline{\alpha}_{12}^* \tag{A.82}$$

with $\underline{\alpha}_{12}^* > 0$.

Still applying the same reasoning, we can establish that when $\gamma \leq \tilde{\gamma}$, there exists a threshold $\overline{\alpha}_{12}^* > 0$ such that

$$v_1(0,0) - v_1(1,0) > 0 \text{ if and only if } \alpha_{12} > \overline{\alpha}_{12}^* \tag{A.83}$$

Hence, we have demonstrated that provided that α_{12} and α_{21} are high enough, if $\gamma \leq \tilde{\gamma}$ or if $\gamma > \tilde{\gamma}$ but $a_{10} > \theta_1$ $(\lambda_1^* = 0, \lambda_2^* = 0)$ is necessarily an equilibrium outcome. This concludes the proof of Theorem 8.

A.2 Market Games

In each period, a seller decides whether or not to sell the good to the buyer.[27] If the seller decides to sell the good to the buyer, she incurs a fixed sunk production cost. For example, this cost could be the cost of job-hunting for a worker supplying her labor. The seller also sets price of the good she is selling. If the seller decides to be active, the buyer decides whether or not to buy the good.[28] Assuming that the good has a fixed value, we can characterize the SPE of this game and establish a result similar to Theorem 4

The value of the good is $V_b > 0$. I assume that the buyer can 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 $a_{s,\infty} = 0$.*

[27] Denoting a_s the action of the seller, $a_s = 1$ if the seller decides to be active and sell her good to the buyer and $a_s = 0$ otherwise.

[28] Denoting a_b the action of the buyer, $a_b = 1$ if the buyer decides to buy the good, and $a_b = 0$ otherwise

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

$$\pi(q_1) = \begin{cases} 1 - q_1 \text{ if } q_1 \in [0,1] \\ 0 \text{ if } q_1 > 1 \text{ and} \\ 1 \text{ otherwise.} \end{cases} \tag{A.84}$$

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

$$p_{bs,1} = \max(0, 1/4 - c - (\alpha_{bs} - \alpha_{ss})/2). \tag{A.85}$$

The punishment is incentive compatible for the buyer if

$$p_{bs,1} < \frac{\gamma}{2} - \max(V_b - 1/2, 0). \tag{A.86}$$

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.

$$\tilde{\gamma}_1 = \max(0, 1/4 - c - (\alpha_{bs} - \alpha_{ss})/2) + \max(V_b - 1/2, 0). \tag{A.87}$$

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,

$$p_{bs,2} = \max(0, 1/4 - c - (\alpha_{bs} + \alpha_{ss})/2) \tag{A.88}$$

and the punishment is incentive-compatible when

$$p_{bs,2} < \frac{\gamma}{2} - \max(V_b - 1/2, 0), \tag{A.89}$$

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

$$\tilde{\gamma}_2 = \max(0, 1/4 - c - (\alpha_{bs} + \alpha_{ss})/2) + \max(V_b - 1/2, 0) \tag{A.90}$$

Comparing (A.87) with (A.90), 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.