Can the Culture of Honor Lead to Inefficient Conventions?

Experimental evidence from India

Benjamin Brooks
Karla Hoff
Priyanka Pandey
Abstract

Experiments in the United States have found that pairs of individuals are generally able to form socially efficient conventions in coordination games of common interest in a remarkably short time. This paper shows that this ability is not universal. The paper reports the results of a field experiment in India in which pairs of men from high and low castes repeatedly played a coordination game of common interest. Low-caste pairs overwhelmingly coordinated on the efficient equilibrium, consistent with earlier findings. In contrast, high-caste pairs coordinated on the efficient equilibrium at a much lower rate, with only 47 percent in efficient coordination in the final period of the experiment. The study traces the divergence in outcomes to how an individual responds to the low payoff he obtains when he attempts efficient coordination but his partner does not. After this event, high-caste men are significantly less likely than low-caste men to continue trying for efficiency. The limited ability to form the efficient convention can be explained by the framing effect of the culture of honor among high-caste men, which may lead them to interpret this event as a challenge to their honor, which triggers a retaliatory response.

This paper is a product of the Macroeconomics and Growth Team, Development Research Group. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://econ.worldbank.org. The authors may be contacted atkhoff@worldbank.org.
Can the Culture of Honor Lead to Inefficient Conventions?
Experimental evidence from India*

Benjamin Brooks        Karla Hoff          Priyanka Pandey

Keywords: Culture, coordination, conventions, framing, Stag Hunt, learning, caste, India
JEL codes: C72, O12, Z13.

*Brooks: Department of Economics, University of Chicago, babrooks@uchicago.edu; Hoff: The World Bank, khoff@worldbank.org; Pandey: The World Bank, ppandey@worldbank.org. We greatly benefitted from early discussions with Paul Attewell, Gary Charness, Ernst Fehr, Rocco Macchiavello, Matthew Rabin, Jennifer Stellar, and Joseph Stiglitz. We thank seminar participants at Indiana University, Princeton, the University of California–San Diego, the University of California–Santa Barbara, Chicago Booth, the University of Memphis, and Warwick University. We also thank Cristina Bicchieri and Allison Demeritt for detailed and helpful comments. Finally, thanks to Sonal Vats for excellent research assistance and the World Bank Research Support Budget for financial support.
1 Introduction

Economists increasingly emphasize the role of conventions in determining the behavior of economic agents. Conventions influence a vast range of behaviors, for example, respect for others’ property rights, voluntary contributions to public goods, and notions of fairness as constraints on profit-seeking. Conventions are the basis for many formal institutions, and the law can overturn established conventions only with great difficulty.\footnote{Examples in which established conventions have persisted contrary to legal prescription include liability rules (Ellickson, 1991), corruption (Basu, 2015; Dixit, 2016), and violence against women.} Many conventions arise endogenously through a process of coordination (Lewis, 1969). Earlier work on coordination games of common interest led to the comforting view that fixed groups of individuals that are sufficiently small will generally converge on efficient conventions (Van Huyck et al., 1990; Knez and Camerer, 1994). Weber (2006) writes that fixed pairs of individuals “are almost assured to coordinate on the efficient equilibrium” and shows that efficient conventions can be maintained if the coordinating group starts small and gradually grows into a larger organization.

This paper tests the hypothesis that small groups of individuals can efficiently coordinate and finds that there are limitations to the positive results obtained by previous studies. Specifically, we study the outcome of a coordination game of common interest played in villages in northern India. This region has a long history of institutional failures that have stymied economic development (Drèze and Gazdar, 1997), which makes it an interesting and important setting in which to study how individuals solve coordination problems. In contrast to the prior literature, we find that whether or not fixed pairs of individuals are able to establish an efficient convention depends on the cultural subgroup, i.e., social caste, to which they belong. In particular, pairs in which the subjects are high caste establish efficient conventions at a much lower rate than do pairs with low-caste subjects. We argue, experimentally and theoretically, that this difference can be explained by the culture of honor that exists among the high castes but is less prevalent among low castes. The culture of honor may make it hard for high-caste subjects to “forgive” episodes of miscoordination that occur while trying to form or sustain an efficient convention. Our findings show that culture can have a strong influence on individuals’ ability to form efficient conventions, and thus create institutions, that are conducive to economic development.

Our field experiment was conducted in a rural part of the north Indian state of Uttar Pradesh. Fixed pairs of men, who were drawn from the extreme top and bottom of the Indian caste hierarchy, repeatedly played the coordination game of common interest known as the Stag Hunt. These high and low castes are hereditary social groups that speak the same
language, overlap in the distributions of income and education, and live in many of the same villages. At the same time, these groups occupy separate hamlets within the villages and practice caste endogamy, which restricts social interaction between the groups and reinforces substantial cultural differences.

In the Stag Hunt, a player must choose between two actions: “hunt Hare,” which yields a modest but riskless payoff, and “hunt Stag,” which yields a high payoff only if the other player chooses Stag as well. Playing Stag when the other player chooses Hare results in a loss, relative to the reference point of an initial endowment, which we term the loser’s payoff. The Stag Hunt has two pure-strategy Nash equilibria, (Stag,Stag) and (Hare,Hare), of which the former Pareto dominates the latter. This game captures the problem of forming a social contract that benefits all members of society (Skyrms, 2003). The problem of forming an efficient convention is essentially the problem of converging to a common understanding that everyone in the game will play Stag.

In our experiment, a player knew that his partner was a man in his village of a given high or low caste status. Pairs were formed anonymously and included the three possible combinations of caste statuses: high caste with high caste (HH), low caste with low caste (LL), and low caste with high caste (LH). A priori, there are many reasonable conjectures for how the subjects’ caste statuses would correlate with the outcome of coordination. One might predict that coordination would be more efficient between individuals with the same caste status, since social proximity might lead to greater trust and concern for the partner’s welfare. One might also predict that education might make it easier to understand and “solve” the game, which would suggest that the generally better-educated high-caste players might coordinate more efficiently.

The striking result of the experiment was that low-caste subjects coordinated far more efficiently than their high-caste counterparts. For example, 73 percent of LL pairs played the efficient equilibrium in the final period of the partnerships, compared to 50 percent of LH pairs, and only 32 percent of HH pairs. Moreover, while LL pairs that started in miscoordination tended to converge to the efficient equilibrium, HH pairs that started in miscoordination tended to converge to the inefficient equilibrium. The LH pairs displayed no clear trend in either direction.

---

2The name Stag Hunt comes from a parable told by Rousseau: Each of several hunters has a choice between hunting a stag and hunting a hare. Regardless of what others do, pursuing the hare is riskless in the sense that it will always result in a meager meal. If everyone pursues the stag, the stag hunt will be successful, and each hunter will obtain a rich meal. However, if some but not all hunters pursue the hare, the stag hunt will be unsuccessful, and a player who attempts to hunt the stag without the cooperation of others will not eat at all.
These differences in pair outcomes resulted from substantial caste differences in individual behavior. Even in the first period of the game, 68 percent of low-caste subjects play *Stag* compared to only 53 percent of high-caste subjects. This difference is modest, however, in comparison with the differences in behavior that emerged in response to the history of play. A natural descriptive statistic to consider is the likelihood that an individual attempts efficient coordination conditional on the previous period’s outcome. For example, if players coordinate on (*Stag*, *Stag*), one might suppose that the players would be very likely to play *Stag* again, thereby sustaining the efficient convention. This is indeed the case: for every pair of caste statuses, almost 90 percent of individuals play *Stag* after efficient coordination. In fact, there is only one outcome after which the caste difference in behavior is statistically significant: the outcome in which a player receives the loser’s payoff. In this event, 72 percent of low-caste players choose *Stag* again, versus 42 percent of high-caste players. This difference is even starker if one compares same-caste pairs: 68 percent of low-caste subjects with low-caste partners will play *Stag* after the loser’s payoff, versus only 32 percent of high-caste subjects with high-caste partners, a difference of 36 percentage points! Caste differences after all other one-period histories are much smaller and statistically insignificant. The same qualitative results obtain in regressions that add proxies for education and wealth.

Why do high- and low-caste men behave so differently after obtaining the loser’s payoff? We hypothesize that this difference is due to the culture of honor that is typical of the high castes of northern India but not of low castes (Khare, 1984; Mandelbaum, 1993). The culture of honor has been well-documented by anthropologists and sociologists and has been observed among groups all over the world and throughout history, including in northern India, Mediterranean societies, Saudi Arabia, many of the traditional societies of Eurasia and Africa, and the southern United States (Schneider, 1971; Nisbett and Cohen, 1996; Gregg, 2005). In the culture of honor, certain behaviors are regarded as transgressive (that is, challenges to honor), and individuals deter transgressions against themselves by maintaining a reputation for aggressive retaliation (defending their honor). As the anthropologist Joyce Pettigrew (1991, p. 174) writes, the culture of honor “enjoined that revenge be exacted for personal insults and damage to person or property.” As we confirm with a vignette-based survey, the concern for honor is much more characteristic of high castes than it is of low castes, perhaps because low castes have traditionally been denied the right to accumulate wealth and achieve status.

In many ways, the outcome (*Stag*, *Hare*) fits the bill for the kind of event that would elicit an aggressive, retaliatory response under the culture of honor. A subject that plays *Stag* is “harmed” by a partner who plays *Hare*, in the sense that the partner’s choice of *Hare* imposes on him a loss. The only “honorable” response may be to punish the partner,
and the only way to harm the partner is by playing *Hare* in the next period to prevent the partner from obtaining the high return from efficient coordination. Of course, an alternative and perfectly valid interpretation of the outcome (*Stag*, *Hare*) is that the partner played *Hare* simply because he believed that the other player was going to choose *Hare* as well. Miscoordination was an unfortunate mistake and not indicative of any malicious intent. Intent is, however, largely besides the point in the culture of honor. Honorable men have a rule of thumb to retaliate against slights, regardless of their motive.

Note that if miscoordination never occurred, no one would get the loser’s payoff, and whether or not individuals have the culture of honor would not affect the outcome of the game. But if the only way for players to form and sustain an efficient convention is along a path of behavior that involves miscoordination, then the loser’s payoff will occur and the culture of honor could affect behavior and steer coordination towards inefficient outcomes. We illustrate this mechanism formally within the learning model of logistic fictitious play (Fudenberg and Levine, 1998). This model posits that individuals play a noisy best response to the historical distribution of others’ actions. We add to that standard framework the assumption that individuals may have the culture of honor—that is, a preference for punishing others who have offended them in the previous period. If a player is offended by getting the loser’s payoff, then he prefers that his partner receive a low payoff in the subsequent period. We show that if this preference for punishment is sufficiently strong, then efficient coordination is unsustainable in the long run. The logic is simple: Even if players initially expect and actually achieve a high degree of efficient coordination, mistakes and miscoordination will inevitably occur due to the noise in behavior. Miscoordination causes one of the players to receive the loser’s payoff, and the subsequent punishment lowers the empirical frequency of playing *Stag*. This in turn leads even players who did not receive the loser’s payoff to play *Stag* less often, which increases the likelihood of miscoordination, leading to more challenges to honor and the unraveling of the efficient convention.

We also consider and find no support for two alternative hypotheses—caste differences in trust and caste differences in self-efficacy. The first hypothesis is that high-caste men, compared to low-caste men, begin the game believing that their partners are less likely to play *Stag*, so that high-caste men are less “trusting” of their partners. To test this hypothesis, we implemented a second field experiment in which high- and low-caste subjects played the following investment game: one player, the principal, chooses whether or not to give his endowment to another player, the agent. If the principal hands over the endowment, it grows by a large amount. The agent then chooses whether or not to transfer money back to the principal. The principal would expect to gain from handing over his endowment only if he trusts the agent to return it, along with a share of the gain. We implemented this
experiment with all combinations of caste statuses in the principal and agent roles. We found no statistically significant caste difference in sharing. In fact, slightly more principals shared when both principal and agent were high caste than when both were low caste.

Another hypothesis that would explain the caste difference in behavior in the Stag Hunt experiment is a low sense of self-efficacy among low-caste men compared to high-caste men. That hypothesis could also explain a pattern that emerged in our vignette-based survey, which is that low-caste men were much less likely than high-caste men to want to respond violently to the transgressions described in the vignettes. We conducted a second survey to test for caste differences in self-efficacy and found no evidence that low-caste men have lower self-efficacy than high-caste men.

This paper contributes to the rapidly growing literature on the interaction between culture and economic behavior (see, for example, Greif, 2006; Bénabou, 2008; Herrmann, Thöni and Gächter, 2008; Hoff and Stiglitz, 2010, 2016; Hoff, Kshetramade and Fehr, 2011; Nunn, 2012; Jackson and Xing, 2014; World Bank, 2015). Our contribution is to show that culture can block the emergence of efficient conventions in a strategic setting in which prior experimental work with US subjects found that the outcome was almost always efficient. To the best of our knowledge, this is the first paper to link culture to the ability to form efficient conventions in coordination games of common interest.

Numerous studies, including those cited above, have shown that people from different cultures will behave differently in otherwise identical situations, and also that behavior can depend on seemingly irrelevant details of context (Kahneman and Tversky, 1984). An explanation that is broadly accepted in behavioral economics, psychology, sociology, and anthropology is that human decision making is carried out using a set of mental tools, which have been variously referred to as schemas and mental models (Swidler, 1986; D’Andrade, 1995; Nisbett and Cohen, 1996; DiMaggio, 1997; Fiske et al., 1998). In brief, humans process information and interpret the world around them by matching their observations to a relatively sparse set of known patterns, which may lead them to select a course of action by automatically invoking a rule of thumb. The particular patterns and rules of thumb that an individual uses are influenced by history and experience. Thus, individuals who share a common experience will tend to process information and interact with the world in a similar manner.

“Culture” in general usage is a broad and amorphous concept, and it has acquired different technical meanings in different disciplines. In our view, any reasonable definition of culture for an economist should include those common and distinct elements of how a group of individuals perceive, interpret, and interact with the wider economy. The culture of honor, in particular, has been described as such a “mediative concept” by the anthropologist
Pitt-Rivers (1966, p. 244) and as “a code for both interpretation and action” by the anthropologist Friedrich (1977, p. 283). The aspect of this mediation on which this paper focuses is that the culture of honor may cause an individual to adversely respond to miscoordination as if it were an offense, even though this event has no inherent meaning beyond reflecting an incongruence between the two players’ expectations of each others’ actions.

This paper also contributes to the understanding of institutional failures in Uttar Pradesh that can be interpreted as the result of inefficient coordination (Drèze and Gazdar, 1997). In their case study of the village of Palanpur, Drèze and Sharma (1998) note the conspicuous absence of collective action, which they view as the cause of the most serious failures of development in the village (p. 67). Villagers failed to collaborate to address common problems in agriculture, drainage, education, and public health. Before the emancipation of the laboring classes that came with Indian independence, the services of street maintenance and sanitation in villages were provided by those in the low ranks of the caste hierarchy via a system of customary obligation. As of fifty years later, “the challenge of creating [institutions]... to address those needs has been largely unmet.” The culture of honor would have aided individuals’ efforts to protect their wealth in the face of frequent invasions in past centuries, but the days of invasion and the absence of formal institutions are long gone in this part of India. Yet, the culture of honor has persisted due to its private benefits, such as improving the marriage prospects of one’s children and therefore increasing the family’s wealth (Derné, 1994, p. 286).

The same culture of honor may have impeded villagers ability to find efficient solutions to the coordination problems described by Dréze and coauthors, as we have found in our experiment.

As we discuss in detail in the conclusion, the mental models that individuals use are not immutable. Psychologists and sociologists view culture as a “dynamic interaction between mind and environment... with mental structures selecting aspects of the environment as salient, and environments selectively reinforcing mental representations” (DiMaggio and Markus, 2010, p. 347). Recent studies in economics demonstrate successful interventions to change the mental models that individuals use in making particular decisions (Bertrand et al., 2004; Blattman et al., 2012; Heller et al., 2015). The link between culture and co-
ordination therefore points to interventions that could promote institutional reform, either by changing the perception of coordination failure as an offense, or by changing the rule of thumb to retaliate against such offenses.

The next section of the paper describes the design of our Stag Hunt experiment. Section 3 analyzes the results. Section 4 presents our survey on attitudes towards punishment. Section 5 develops a theoretical mechanism by which the culture of honor can impede efficient coordination. Section 6 addresses alternative hypotheses, and Section 7 concludes. Appendices contain the formal analysis of the model of Section 5 and supporting materials from the experiments.

2 The Stag Hunt Experiment

2.1 Subjects

In September 2005, we conducted a field experiment in the northern Indian state of Uttar Pradesh. We recruited 122 male subjects from 10 villages in Bakshi Ka Talab in the district of Lucknow. In each village, we recruited an equal number (generally 6) of high- and low-caste subjects using systematic sampling that covered the entire village. This was done with the help of one to three village residents who were knowledgeable about the location of high- and low-caste hamlets. As an example, if the village had only one low-caste hamlet and if thirty households resided there, then every fifth household was selected.

After explaining to the household that we were using a game to study how people behave, we asked if one man in the household would volunteer to take part in the game, in which he would earn money. In every selected household, a person volunteered. The caste composition of the high-caste subjects was: Brahmin (55 percent), Thakur (44), and Lala (1). Among the low-caste subjects, the composition was: Chamar (56), Rawat (38), Dhanuk (3), and Pasi (2). Two percent of the low-caste sample identified themselves by the generic term for low caste, harijan.

2.2 Procedures

We held only one session in each village so that no subject would have prior knowledge of the game. In each session, the low- and high-caste subjects were kept in separate sites, usually in their neighborhoods, and never visible to one another. At the time of the game, no player knew who the players were at the other location. Cell phone service was unreliable, so runners were used to communicate to each player his partner’s actions as the experiment progressed.
The same team of experimenters handled all sessions. Most team members were university students from Lucknow. At the beginning of the session, the team leader explained the game first in one site and then in the other site. Thus, all of the subjects received their instructions from the same person. Each subject was assigned a monitor, who stayed with the subject until the experiment was over. Before the experiment began, the monitor reviewed the instructions and tested the subject’s understanding. Two subjects were dismissed because they did not understand the game. A subject was not allowed to talk to anyone except his monitor for the duration of the experiment. Each session lasted about four hours. In post-play interviews, the subjects were asked about their landholdings, education, and type of housing.

2.3 Experimental Design

The subjects were organized into pairs and asked to play ten periods of the Stag Hunt, depicted in Figure 1. In each period, every subject was given an endowment of 6 blue tokens, representing one rupee each (so that the 6 tokens were equivalent to 15 US cents). Each subject was then asked to choose between contributing 2 or 6 tokens to a common pool. If a player contributed 2, he would receive back 3 tokens, regardless of the choice of the other player, for a gain of 1 relative to the endowment. If a player contributed the entire endowment of 6, he would receive a payoff of 10 if the other player also contributed 6, for a gain of 4 relative to the endowment, but he would receive only 3 if the partner contributed 2. In the latter case, the player lost half of his endowment while his partner gained 1 token. We refer to this loss of 3 tokens as the loser’s payoff. The device of presenting the players with an endowment was done to frame the loser’s payoff as a loss in the sense of Kahneman and Tversky (1984).

A subject was never told the identity of his partner, but only that he was a man in his village of either “General” (high) or “Scheduled” (low) caste. Each subject played the Stag Hunt for five periods with a partner of high caste status and for five periods with a partner of low caste status. We used a counterbalanced design to avoid confounding the caste status of the pair and the order of play: In Cohort I, subjects were organized into \( LL \) and \( HH \) pairs for periods 1-5, followed by \( LH \) for periods 6-10. In Cohort II, the order was reversed. Players were randomly assigned to a cohort.

To make salient the history of play in a fixed pairing, each monitor gave his subject a “game box” in period 1 and a second game box in period 6. Each box was divided into five columns, corresponding to the five periods of a fixed pairing. Each column was additionally divided into three rows. In the first period, a player indicated his strategy by placing either
6 or 2 of his blue tokens into the bottom row of the first column. If he placed only 2 tokens, then he put the remaining 4 into his personal envelope. At the end of the first period, a messenger came and put orange tokens into the top row of the first column to indicate the partner’s action (either 6 or 2). The monitor then put blue tokens into the middle row of the first column to show the player his earnings from the period (either 10 or 3). Play then continued into the second period using the same format with the second column, and so on. The messengers never spoke to the players, and the monitor was instructed not to look inside the box (which had an opaque cover) when the player or messenger opened it, so that the monitor would never know the player’s actions and could not indicate approval or disapproval.

In the remainder of the paper, we will use the label “Stag” to refer to the action “Contribute 6” and the label “Hare” to refer to “Contribute 2.” We wish to emphasize, however, that the experiment was explained to the subjects using neutral language: invest 6 tokens or invest 2 tokens in the common pool.

After the tenth period, players received their payoffs in private and in cash. Each experimental session lasted about four hours. Mean earnings were 77 rupees, which was 77 percent of the maximum of 100 rupees and approximately 1.5 times the daily unskilled wage.

## 3 Results of the Stag Hunt

### 3.1 Pair Outcomes

We will now describe the outcome of the Stag Hunt experiment, starting with results at the level of pairs. In Figure 2, the top panel shows the distribution of outcomes by type of pair over all periods of a pairing, and the bottom panel shows the distribution of outcomes in
the final period of a pairing, i.e., periods 5 and 10. A striking pattern is that pairs with low-caste players achieved much more efficient outcomes than pairs with high-caste players. Over the course of the entire experiment, LL pairs spent 59 percent of periods in efficient coordination (Stag,Stag), versus 49 percent for LH pairs, and only 34 percent for HH pairs. The contrast is even starker for periods 5 and 10, after the fixed pairs have had ample time to form a convention. In these final periods of a pairing, the proportion of pairs playing (Stag,Stag) was 73 percent for LL, 50 percent for LH, and only 32 percent for HH. All of these differences are statistically significant at the 5 percent level.\footnote{All standard errors for the pair analysis are clustered by village.}

In addition, HH pairs spend much more time in inefficient coordination on (Hare,Hare) than do LH or LL. In particular, HH pairs spent 29 percent of all periods and 35 percent of final periods in inefficient coordination. The comparable numbers for LH and LL pairs pooled are 11 percent and 9 percent. We reject the hypothesis that LL and LH pairs were in inefficient coordination at the same frequencies as HH at the 5 percent level of significance, both for all periods pooled and for final periods.

Figure 3 shows the evolution of behavior across different pair types. In the first five periods, the proportion of LL pairs with the outcome (Stag,Stag) increased from 27 to 67 percent, a statistically significant change ($p < 0.001$). In contrast, there was no statistically significant time trend in the proportion of efficient outcomes among HH and LH pairs, and
a plurality of both pair types were in miscoordination or in the inefficient equilibrium at period 5. This pattern is strikingly different from the finding of Van Huyck, Battalio and Beil (1990) that the frequency of efficient coordination in a Stag Hunt increased over the first five periods and then stayed at that high level for the remaining five periods of a ten-period game.

For periods 6 to 10, LL pairs consistently maintained an efficient convention, with between 67 and 80 percent of pairs playing (Stag, Stag). HH pairs initially started without a single pair at the inefficient equilibrium, though a majority started in miscoordination. However, the fraction of pairs playing (Hare, Hare) rose from zero in the initial period of the pairing to 33 percent in the last two periods, whereas there was no time trend to the proportion playing (Stag, Stag). Evidently, the HH pairs learned to avoid disequilibrium by largely settling on an inefficient convention.\textsuperscript{6} Compared to HH pairs, the time trends for LH are more erratic. A majority achieved efficiency in period 10, though this share is not significantly different from 50 percent at the 5 percent level.

To sum up, high caste status is associated with less efficient coordination. Most LL pairs established an efficient convention, while most HH pairs did not. Outcomes for the LH pairs were in between those of LL and HH. The evidence also suggests that social distance was not

\textsuperscript{6}We know that the H players who move to (Hare, Hare) are generally moving from (Stag, Hare) or (Hare, Stag), since almost 90 percent of H players in (Stag, Stag) in a given period of a fixed pairing continue to play Stag in the next period. Moreover, as indicated in Table 1 below, once the outcome (Hare, Hare) is reached, 74 percent of subjects in HH continue to play Hare in the subsequent period.
an obstacle to efficient coordination. Mixed-status pairs achieved a higher rate of efficient coordination and a substantially lower rate of inefficient coordination than $HH$ pairs.

### 3.2 Individual Behavior

To understand the pair outcomes, we need to unpack how individual behavior varied by caste. Table 1 shows the proportion of subjects playing Stag, broken down by caste status and pair type. Column 1 reports the proportions for all periods. Column 2 reports the proportions in the initial period of a fixed pairing, i.e., periods 1 and 6. Columns 3 to 7 report the proportions conditional on the outcome in the preceding period. We find these statistics useful for understanding how players used the history of the game to determine their actions. In principle, the entire history could factor into the decision, and the mapping from histories to actions could be quite complicated. Conditioning on the entire history is both impractical and unlikely to be edifying. The previous period’s outcome, on the other hand, is in all likelihood the most prominent feature of the history and therefore might capture, to a first order, the relationship between history and behavior.

As column 2 shows, there is a caste gap in behavior even in the initial periods of a pairing, before there is any history with that player upon which to condition. 68 percent of low-caste subjects play Stag in the initial period, compared to only 53 percent of high-caste subjects, a difference that is significant at the 5 percent level according to a two-sided $t$-test. This gap is similar across single- and mixed-status pairs.

Columns 3 to 7 show the proportion who played Stag conditional on the previous period’s outcome $(x,y)$, where $x$ denotes the player’s own action and $y$ is the partner’s action. For instance, column 3 shows that after the efficient equilibrium outcome ($Stag, Stag$), subjects in all three pair types played Stag at virtually the same high rate—nearly 90 percent. For other histories, however, there are substantial caste differences in behavior. The largest difference occurs after the outcome $(Stag, Hare)$, that is, after the subject received the loser’s payoff. After this history, 71 percent of low-caste players chose Stag in the next period compared to only 42 percent of high-caste subjects, a difference of 29 percentage points! This is the only history after which the caste difference is statistically significant according to a two-sided $t$-test ($p < 0.05$). The gap is even more stark if one compares the single-caste pairs, in which 36 percentage points more low-caste than high-caste players tried again for efficiency.

It is important to note that there are also substantial caste gaps after the histories $(Hare, Stag)$ and especially $(Hare, Hare)$. In the latter case, 21 percentage points more low-caste subjects than high-caste subjects played Stag, and for single-caste pairs the gap was 24
percentage points. These differences contributed to the divergence in pair outcomes, though the data indicate the particular importance of the last-period outcome (Stag, Hare).

Caste status is strongly correlated with wealth and education. We next consider whether these covariates explain the variation in behavior by caste status. We collected data in post-play interviews on land ownership, education, and housing. Figure 4 shows, by caste status, the distribution of land owned by the households represented in our sample, as well as the proportions of subjects who completed high school and live in houses constructed only of mud (with thatch or tiled roofs). High-caste subjects own more land, are more likely to have a high school education, and are more likely to live in a house that is not made out of mud. However, there is substantial within-group variation, and the distributions of these characteristics overlap between high and low castes. For example, approximately 15 percent of the low-caste sample completed high school, compared to 18 of the high castes; 63 percent of low-caste subjects live in a mud house, compared to 19 of high-caste subjects.

Table 1: Proportion of players who chose Stag.

|       | All periods | Initial periods | (Stag, Stag) | (Stag, Hare) | (Hare, Stag) | (Hare, Hare) |
|-------|-------------|-----------------|--------------|--------------|--------------|--------------|
| \(L\) | 0.74        | 0.68            | 0.88         | 0.71         | 0.58         | 0.52         |
| in \(LL\) | 0.74 | 0.65 | 0.86 | 0.68 | 0.66 | 0.50 |
| in \(LH\) | 0.74 | 0.70 | 0.89 | 0.72 | 0.47 | 0.53 |
| \(H\) | 0.58        | 0.53            | 0.87         | 0.42         | 0.46         | 0.31         |
| in \(LH\) | 0.64 | 0.55 | 0.86 | 0.56 | 0.46 | 0.40 |
| in \(HH\) | 0.53 | 0.52 | 0.88 | 0.32 | 0.47 | 0.26 |
| \(L - H\) | 0.16        | 0.15            | 0.01         | 0.29         | 0.12         | 0.21         |
| \(LL - HH\) | 0.21 | 0.13 | -0.02 | 0.36 | 0.19 | 0.24 |
| \(N\) | 1,210       | 242             | 452          | 181          | 181          | 154          |

\(7\) Houses were classified as either brick (\textit{pucca}), impermanent materials like sticks and mud, or a mixture of the two.

\(8\) Using data from the 1997-98 Survey of Living Conditions in Uttar Pradesh, we find that land ownership, high-school education, and housing type together explain between 30 and 40 percent of the variation in consumption for both high and low castes (Hoff, Kshetramade and Fehr, 2011).
Figure 4: Distribution of covariates by caste status. The left panel gives the cumulative distribution of the logarithm of land holdings in hectares. The right panel depicts the proportion of subjects who have a high school education and who live in a mud house, respectively, by caste status.

We use probit regressions to test whether or not differences by caste status in the probability of playing Stag can be explained by our proxies for wealth and education. Probit regressions will also allow us to control for unobserved individual characteristics. In particular, some individuals may be more trusting or have more pro-social preferences, which would be correlated with efficient outcomes. We proxy for such characteristics using the player’s first-period action, which we refer to as that player’s “type.”

The results of probit regressions are displayed in Tables 2 and 3. For each variable, we report coefficients and clustered standard errors for average marginal effects. The omitted case in regressions in Table 2 is a low-caste player. For Table 3, the omitted case is a low-caste player in an LL pair.

Column 1 of Tables 2 and 3 shows that there is a modest and marginally significant difference in behavior between low- and high-caste individuals in the initial period: high caste status is associated with a 16 percentage point lower likelihood of playing Stag. By far the largest difference in behavior occurs after subjects received the loser’s payoff, which is reported in column 3. In this case, high caste status is associated with a decrease of 36 percentage points in the likelihood of playing Stag in the subsequent period. The gap grows substantially if one compares a subject in LL versus one in HH: Being in an HH pair is associated with a 42 percentage point lower likelihood of playing Stag. These results show that controlling for covariates and the initial propensity to play Stag results in an even larger estimate of the effect of the player’s caste status on the propensity to play Stag. Differences after other one-period histories are smaller and statistically insignificant. In particular, for every history except (Stag, Hare), a t-test fails to reject the hypothesis that the coefficients

---

9In the regressions, type is 1 if the player played Stag in period 1 and the period is greater than 1, and 0 otherwise.

10Players were organized into fixed pairs in groups of four, such that individuals A, B, C, and D were assigned to pairs (A, B) and (C, D) for periods 1–5 and to pairs (A, C) and (B, D) for periods 6–10. Standard errors are clustered by these four-tuples of players.
Table 2: Probit regression analysis for $H$ versus $L$. Marginal probabilities that a player chose Stag are shown, with standard errors of the marginal effects in parentheses.

|               | Initial | $(Stag,Stag)$ | $(Stag,Hare)$ | $(Hare,Stag)$ | $(Hare,Hare)$ |
|---------------|---------|---------------|---------------|---------------|---------------|
| $H$           | -0.161* | 0.0169        | -0.363***     | -0.0931       | -0.226        |
|               | (0.0712)| (0.0362)      | (0.100)       | (0.114)       | (0.170)       |
| Type          | 0.347***| 0.110**       | -0.0509       | 0.194         | -0.118        |
|               | (0.0573)| (0.0452)      | (0.0612)      | (0.109)       | (0.105)       |
| Land          | 0.00579 | -0.00138      | 0.00642       | -0.00156      | -0.00505      |
|               | (0.00399)| (0.00239)    | (0.00437)     | (0.00686)     | (0.00770)     |
| High school   | 0.0643  | -0.0779       | 0.160         | -0.0921       | -0.140        |
|               | (0.0870)| (0.0587)      | (0.102)       | (0.137)       | (0.144)       |
| Non-mud house | 0.0523  | 0.0247        | 0.0559        | 0.0506        | 0.0309        |
|               | (0.0628)| (0.0447)      | (0.0855)      | (0.102)       | (0.115)       |

| $N$        | 242  | 452  | 181  | 181  | 154  |

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

4 Understanding Differences in Caste Culture

What explains the difference in behavior between high and low castes that occurs after the loser’s payoff? We present empirical evidence in this section and a theoretical model in the next section to support the hypothesis that a stronger culture of honor among the high castes than among the low castes is driving the difference in behavior.

As noted above, a man who adheres to the culture of honor is constantly on guard against affronts that could be construed by others as disrespect. Honor is one of his most valuable possessions. Since the loser’s payoff creates a loss in absolute and in relative terms,
Table 3: Probit regression for $H$ versus $L$ players, by pair status. Marginal probabilities that a player chose Stag are shown, with standard errors of the marginal effects in parentheses.

|                      | Initial | $(\text{Stag, Stag})$ | $(\text{Stag, Hare})$ | $(\text{Hare, Stag})$ | $(\text{Hare, Hare})$ |
|----------------------|---------|------------------------|------------------------|------------------------|------------------------|
| $HH$                 | -0.133  | 0.0360                 | -0.423**               | -0.159                 | -0.254                 |
|                      | (0.0869)| (0.0362)               | (0.119)                | (0.126)                | (0.201)                |
| $H$ in $LH$          | -0.0943 | 0.0152                 | -0.184                 | -0.146                 | -0.121                 |
|                      | (0.102) | (0.0576)               | (0.131)                | (0.132)                | (0.215)                |
| $L$ in $LH$          | 0.0974  | 0.0152                 | 0.0741                 | -0.130                 | 0.0114                 |
|                      | (0.0970)| (0.0386)               | (0.122)                | (0.133)                | (0.173)                |
| Type                 | 0.354***| 0.109**                | -0.0516                | 0.174                  | -0.122                 |
|                      | (0.0546)| (0.0439)               | (0.0645)               | (0.118)                | (0.105)                |
| Land                 | 0.00584 | -0.00153               | 0.00541                | -0.00156               | -0.00514               |
|                      | (0.00400)| (0.00242)               | (0.00480)               | (0.00682)               | (0.00775)               |
| High school          | 0.0622  | -0.0777                | 0.183                  | -0.0840                | -0.124                 |
|                      | (0.0879)| (0.0593)               | (0.0981)               | (0.138)                | (0.149)                |
| Non-mud house        | 0.0519  | 0.0234                 | 0.0678                 | 0.0459                 | 0.0358                 |
|                      | (0.0629)| (0.0474)               | (0.0842)               | (0.104)                | (0.112)                |

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

It would be natural for a person who adheres to the culture of honor to interpret the loser’s payoff as an insult to his honor and to retaliate. As shown in the anthropology and sociology literatures, the culture of honor is much stronger among high-caste men than among low-caste men. To test that this is indeed true of the communities from which we drew our samples, we implemented in 2013 a vignette-based survey of the norms for retaliation. The survey was conducted in 22 hamlets in the same area in which we had conducted the Stag Hunt experiment. The design follows a format developed by social psychologists for measuring moral attitudes (see, for example, Haidt, Koller and Dias, 1993). Subjects were presented with vignettes in which one individual, named Dinesh, behaves in a way that harms a second individual, Mahesh, who then responds violently. The subjects were asked whether or not they thought that Mahesh’s response was justified, and they were also asked the open ended-
question of what would they have done in Mahesh’s place. We note that the names Dinesh and Mahesh are not associated with a particular caste.

The first column of Table 4 summarizes the vignettes; the online appendix presents them in full. In the control vignette V0, Dinesh is a thief who robs Mahesh’s house. Mahesh later beats the thief and reports him to the police, who put Dinesh in jail. In vignettes V1–V3, Dinesh harms Mahesh in ways that are not unambiguous violations of law. In vignette V4, the “harm” is inter-caste marriage, with one version of the vignette used for high-caste respondents and another version for low-caste respondents. Each subject was presented with two vignettes. The first was chosen randomly from V0–V3, and the second one was the inter-caste marriage vignette (V4). Our goal in this design was to prevent cross-contamination in the response and also to leave to last the vignette on inter-caste marriage, which upset some respondents. We recruited 121 high-caste men and 120 low-caste men in the same manner and from the same region of Uttar Pradesh in which we had conducted the Stag Hunt experiment.

We expected both high- and low-caste men to respond affirmatively for V0, since Mahesh was in clear violation of the law and the main punishment was obtained through the legal process. On the other hand, since the culture of honor makes individuals more likely to perceive a harm as an “insult to honor” and to retaliate, we expected more high-caste men than low-caste men to think the violent response was justified for V1–V4. This is exactly what we find. For V0, the distribution of responses is not statistically different between high and low castes. For V1–V4, however, a one-sided t-test rejects the null hypothesis that more low-caste than high-caste men think the response was justified, with p < 0.05 for V1–V2 and p < 0.001 for V3–V4. For V4, over 70 percent of high-caste respondents said the violent response was justified, compared to only 22 percent of low-caste respondents.

The surveyor asked each respondent an open-ended question about what he would have done if he were the wronged party. Compared to low-caste men, high-caste men were more likely to mention an aggressive response. Low-caste respondents were more likely to think about how to deescalate a conflict than how to aggravate it and were more likely than high-caste men to offer a conciliatory response. For V1-V3, 37 percent of high-caste men stated that they would do something aggressive, compared to 17 percent of low-caste men. For V4, 78 percent of high-caste men said they would respond with aggression compared to 35 percent of low-caste men.

---

11 For high-caste men, Mahesh Bania marries Dinesh Thakur’s daughter. Bania and Thakur are both high status, but the latter is slightly higher rank. For low-caste respondents, Bania and Thakur were changed to the low-status castes Pasi and Chamar, respectively, with Chamar being slightly higher in status.

12 Respondents were drawn from the same block and district near Lucknow. Half of the hamlets sampled for the survey are also represented in the Stag Hunt sample.
| Description                                                                 | Response was justified | N   |
|----------------------------------------------------------------------------|------------------------|-----|
| V0 Dinesh robs Mahesh’s house, after which Mahesh beats Dinesh, reports him to the police, and Dinesh receives a jail sentence. | 0.90 0.93 0.03 | 61  |
| V1 Dinesh digs a canal through Mahesh’s field without permission. Mahesh later argues with Dinesh in the village and beats him up. | 0.03 0.38 0.35* | 59  |
| V2 Dinesh lets his cattle graze on Mahesh’s fields without permission. Mahesh later argues with Dinesh in the village and beats him up. | 0.20 0.43 0.23* | 60  |
| V3 Dinesh calls Mahesh names and roughs him up. They later meet in public, argue, and Mahesh beats up Dinesh. | 0.10 0.33 0.23*** | 59  |
| V4 Dinesh’s daughter marries Mahesh, even though Mahesh is lower caste than the daughter. Dinesh later beats up Mahesh. | 0.22 0.70 0.48*** | 203 |

* p < 0.05, ** p < 0.01, *** p < 0.001

Table 4: Proportion of respondents who agree that the response was justified in the punishment survey. Standard errors have been clustered by village. Of the 241 respondents, 38 were not able to answer for vignette V4, one was not able to answer for V1, and one was unable to answer for V3. Among the men who did respond for V4, 99 were low-caste and 104 were high-caste.

The caste difference in the proportion who would respond with aggression is in every case statistically significant (p < 0.01). A few examples of aggressive responses are:

“I would do the same as Mahesh because I lost my honor.”
“I would do tit for tat, otherwise people will think I am weak.”
“I would do the same, it is wrong to cause a loss.”
“I would do the same as Mahesh because a mistake is a mistake.”

For the low-caste respondents, non-aggression was much more common. 78 percent of responses for V1-V3 and 63 percent of responses for V4 indicated a non-aggressive approach, and suggested intent to solve the problem, forgive, reconcile, or let things be. Examples of non-aggressive responses were:

“I would talk and find an agreement.”
“I would deal with it peacefully.”
“I would do nothing, tolerate it.”
“I would forgive.”
In sum, high-caste men were much more likely than low-caste men to agree that an offense warrants severe punishment, even when the offender’s intentions were ambiguous. They were also much less likely than low-caste men to suggest a conciliatory response. In some situations, high-caste men explicitly linked their action to the preservation of honor (izzat).

5 A Simple Model of Convention Formation with the Culture of Honor

The preceding section gave evidence that the culture of honor is more common among high-caste than among low-caste men, which may lead high-caste men to play Stag less frequently than low-caste men after receiving the loser’s payoff. What remains to do is to connect the difference in behavior after the loser’s payoff to pairs’ ability to form and sustain an efficient convention. This section develops such a theoretical connection, based on the well-known learning model of logistic fictitious play (Fudenberg and Levine, 1998).

In this model, an individual uses the history of his partner’s actions to forecast future behavior. For example, if a player’s partner has played Stag frequently in the past, then the player expects the partner to play Stag frequently in the future as well. This is admittedly a gross simplification of the real player’s thought process, but one that is likely to capture a first-order effect of the game’s history on players’ beliefs. As the game progresses and players collect more observations, they refine their forecasts of what their partners will do. In a given period, players’ actions are noisy best responses given their current beliefs.

To this basic setup, we add one new feature: Players who adhere to the culture of honor interpret the loser’s payoff as an insult rather than a coordination failure or accident, and insulted players prefer that their partners receive low payoffs in the following period. The addition of these simple state-dependent preferences turns out to have a large impact on which conventions can be learned. If the preference for punishment is sufficiently strong, it will be impossible to learn and sustain the efficient convention.

Here is the formal model. Individuals 1 and 2 repeatedly play the Stag Hunt for periods \( t = 1, 2, \ldots \) In each period, player \( i \in \{1, 2\} \) is in one of two states—normal \((N)\) or insulted \((I)\)—denoted by \( s_{i,t} \). Players are in state \( N \) unless they received the loser’s payoff in the preceding period, in which case they are in state \( I \). In state \( N \), player \( i \)'s utility is just his monetary payoff \( \pi_i \) given in Figure 1. In state \( I \), the player believes that his partner has “insulted” him, and his utility changes to \( \pi_i - \sigma_i \pi_j \), with \( \sigma_i \geq 0 \). If \( \sigma_i = 0 \), being insulted has no effect on preferences, but if \( \sigma_i > 0 \), an insulted player receives utility from his
own monetary payoffs and disutility from his partner’s monetary payoffs.\textsuperscript{13} The preference parameter $\sigma_i$ captures the extent to which the player subscribes to the culture of honor.

Player $i$ in period $t$ takes an action $a_{i,t} \in \{\text{Stag, Hare}\}$. We assume that this action is an approximate best response given player $i$’s beliefs about player $j$’s action. If we let $[a_{j,\tau} = \text{Stag}] = 1$ if player $j$ played Stag in period $\tau$, and 0 otherwise, then at time $t$, player $i$ believes that $j$ will play Stag with probability

$$p_{i,t} = \frac{1}{t} \left( p_{i,1} + \sum_{\tau=1}^{t-1} [a_{j,\tau} = \text{Stag}] \right),$$

where $p_{i,1}$ is some fixed initial belief. Equivalently, $p_{i,t}$ evolves over time according to

$$p_{i,t} - p_{i,t-1} = \frac{1}{t} \left( [a_{j,t-1} = \text{Stag}] - p_{i,t-1} \right). \quad (1)$$

Intuitively, these formulae say that a player believes his partner’s action to be distributed according to the historical empirical distribution of play.

Player $i$ plays a noisy best response given his belief about his partner’s action. Specifically, he plays Stag with probability $b_i(p|s)$ when his state is $s$ and his forecast is $p$:

$$b_i(p|s) = \frac{\exp (u_i(\text{Stag}|p, s))}{\exp (u_i(\text{Stag}|p, s)) + \exp (u_i(\text{Hare}|p, s))}, \quad (2)$$

where $u_i(a|p, s) = p u_i(a, \text{Stag}|s) + (1 - p)u_i(a, \text{Hare}|s)$ is a player’s expected utility from the action $a$ as given in Figure 1, e.g., $u_i(\text{Stag}, \text{Stag}|s) = 4$. The function $b_i$ captures in reduced form the idea that while there is noise in behavior, individuals are more likely to take the action that gives greater expected utility. Since being in the insulted state reduces the expected utility from playing Stag, it follows immediately that $b_i(p|\text{N}) > b_i(p|\text{I})$, i.e., Stag is played with lower probability when insulted than when not insulted.

We are interested in how the culture of honor impacts players’ ability to learn and sustain an efficient convention. In the model, this question can be expressed as: how does the value of $\sigma$ affect the long-run probability of the outcome $(\text{Stag, Stag})$? The Appendix contains a formal analysis, and here we will summarize the main results and intuitions. The core finding is that as $\sigma_i$ increases for either player, the long-run probability of $(\text{Stag, Stag})$ must decrease. For sufficiently large $\sigma$, the only long-run equilibrium will involve a high probability of $(\text{Hare, Hare})$. This means that if the culture of honor is sufficiently important, then efficient behavior cannot be sustained.

\textsuperscript{13}This corresponds to spiteful preferences in the sense of Fehr and Schmidt (1999) and Charness and Rabin (2002).
Let us describe this logic in more detail. Suppose players are in a long-run steady state in which beliefs are close to constant at \((p_i, p_j)\). Then it would have to be the case that beliefs about others’ actions are very close to the empirical distribution that is generated by best responses. Since players with \(\sigma_i > 0\) play *Stag* with lower probability when insulted, the average probability of playing *Stag* depends on the relative frequency of states \(I\) and \(N\). Let \(\Gamma(s_i, s_j|p_i, p_j)\) denote the long-run distribution of \((s_i, s_j)\) when beliefs are fixed at \((p_i, p_j)\). This distribution of states can be used to calculate the average probabilities of playing *Stag*, given beliefs and states. Also let \(b_i(p_i, p_j)\) denote the probability with which player \(i\) chooses *Stag* when beliefs are fixed at \((p_i, p_j)\), averaged across states:\(^{14}\)

\[
\begin{align*}
\Gamma(I, N|p_i, p_j) b_i(p_i|I) + (1 - \Gamma(I, N|p_i, p_j)) b_i(p_i|N).
\end{align*}
\]

(3)

In a steady state in which beliefs are not changing, it must be that beliefs are equal to the long-run empirical average. In other words,

\[
p_i = b_j(p_j, p_i) \quad \forall i = 1, 2, j \neq i.
\]

(4)

A belief that satisfies this condition is called *consistent*, and a *learning equilibrium* is a pair \((p_1, p_2)\) such that both players’ beliefs are consistent.

We argue in the Appendix that for each \(p_j\), there exists a unique \(p_i\) that satisfies (4), which we denote by \(p^*_i(p_j)\). The learning equilibria lie at the intersection of \(p^*_i\) and \((p^*_j)^{-1}\). Moreover, these consistent belief functions \(p^*_i\) are decreasing in both \(\sigma_i\) and \(\sigma_j\). The reason is that as the culture of honor becomes stronger, the probability of playing *Stag* when insulted decreases. Starting from a regime in which *Stag* is played with high probability, a higher \(\sigma_i\) or \(\sigma_j\) increases the likelihood of insults, which depresses the probability of playing *Stag*.

We have been thinking about the behavior induced by a fixed set of beliefs, but in fact beliefs evolve over time according to (1). The literature on stochastic approximation has shown that in the long run, beliefs trend towards an average probability of playing *Stag* that is precisely (3). In fact, as \(t\) becomes large, beliefs approximately evolve according to the dynamical system

\[
\frac{dp_i}{dt} = b_j(p_j, p_i) - p_i.
\]

(5)

\(^{14}\)Since a player enters state \(I\) only after receiving the loser’s payoff, remains in that state for only one period, and only one player can receive the loser’s payoff at a time, it is impossible for both players to be in state \(I\).
Note that in a learning equilibrium, \( \frac{dp_i}{dt} = 0 \) for all \( i \), so that beliefs are not changing. In principle, there can be trajectories that do not converge to a learning equilibrium, but for our game, beliefs will in fact converge to a learning equilibrium. Generically, there are either one or three learning equilibria, the highest and lowest of which are stable, in the sense that beliefs will tend to move towards the equilibrium whenever they are nearby.

Figure 5 illustrates the consistent belief functions for \( \sigma_1, \sigma_2 \in \{0, 0.3, 0.6\} \). The arrows indicate the direction in which the system (5) is trending. For example, the arrows are pointing into the lowest and highest crossings of the consistent belief functions, indicating that they are the stable equilibria. For the cases where there are two stable equilibria, we have drawn the boundary of the basins of attraction of the two equilibria.

In the left-most panel, neither player adheres to the culture of honor, that is, \( \sigma_1 = \sigma_2 = 0 \). In this case, there are two stable learning equilibria, one of which involves a high probability of playing \( Stag \) (\( p_i \approx 0.96 \)). In the center panels, only player 2 adheres to the culture of honor. This means that player 1’s consistent belief is lower, since player 2 plays \( Stag \) less often when he is insulted. For \( \sigma_2 = 0.3 \), there still exists a stable equilibrium in which \( Stag \) is played with high probability of about 0.90, as depicted in center-top. As \( \sigma_2 \) increases further, the consistent belief functions decrease, so that at \( \sigma_2 = 0.6 \) (center-bottom) there is only a single learning equilibrium, in which \( Stag \) is played with probability 0.02. This outcome is even worse than the \( worst \) stable equilibrium when \( \sigma_1 = \sigma_2 = 0 \). The right panels show that when both players adhere to the culture of honor, for \( \sigma = 0.3 \) or \( \sigma = 0.6 \), the only equilibria are inefficient.

We define an efficient learning equilibrium to be one in which the probability of playing \( Stag \) is at least the probability of playing \( Stag \) in the worst stable equilibrium with \( \sigma_1 = \sigma_2 = 0 \). This is a modest criterion for efficient behavior, since the lowest stable probability of \( Stag \) when neither player has the culture of honor is less than 0.021. Figure 6 depicts the set of \( (\sigma_1, \sigma_2) \) pairs in \([0, 1] \times [0, 1]\) for which an efficient equilibrium exists, which is only when the culture of honor is relatively weak for both players. If the culture of honor is sufficiently strong for either player, the only learning equilibria involve a high degree of inefficient coordination.

In sum, the model shows that the outcome of the Stag Hunt experiment can be explained by state-dependent preferences of the high-caste players. Such preferences would cause lower probabilities of playing \( Stag \) after receiving the loser’s payoff, which in turn affects the long-run probabilities of the outcome \((Stag, Stag)\). If low-caste players are always in the normal state, then they may be able to sustain high probabilities of \((Stag, Stag)\) in the long run. On the other hand, if high-caste players are sometimes insulted and therefore have a preference
for punishing their partner after getting the loser’s payoff, then they may be unable to sustain a high probability of playing *Stag*.

The novel ingredient in this model is that players’ preferences depend on how they describe the outcomes to themselves, that is, on how they interpret them. The culture of honor transforms the game by framing the loser’s payoff as a challenge to “honor,” which may evoke in the players an ethic of retaliation and revenge. In some repeated games (most famously the Prisoner’s Dilemma), such retaliation could increase welfare, since it deters deviations that are individually profitable but socially inefficient. In contrast, such retaliation is purely destructive in the Stag Hunt: As long as a player believes his partner will play *Stag*, it is in that player’s self-interest to play *Stag* as well. When players employ learning dynamics such as those described above, retaliation undermines that belief and so undermines the ability to learn and sustain the efficient convention. Retaliation to restore “honor” in the Stag Hunt experiment is not “specific deterrence”—punishment targeted to an action to prevent the partner from repeating it. Nor is it “general deterrence”—punishment that deters third parties from inflicting harms, since the players in the experiment are anony-

---

15 The idea of explicitly modeling how agents construe available alternatives has some precedent in game theory. See, for example, Bacharach (1993) and Binmore and Samuelson (1994).
mous. Hence, players in the Stag Hunt would be unambiguously better off if they could be dissuaded from retaliating after the loser’s payoff.

6 Alternative Hypotheses

6.1 A Difference in Trust?

We have argued that the outcome of the Stag Hunt experiment can be explained by state-dependent preferences of high-caste men that are shaped by the culture of honor. High- and low-caste men have different mental models and different rules of thumb for punishing an offense. An alternative explanation of the lower rate of efficient coordination among high-caste players is a caste difference in how much players trust their partners to take the collaborative action. Some of the results in Table 1 are consistent with the hypothesis that high-caste subjects are less trusting than their low-caste counterparts. In period 1, before there is any history of play to influence decisions, 57 percent of low-caste players chose Stag, compared to only 37 percent of high-caste players.\footnote{Note that the numbers in column 2 of Table 1 pool periods 1 and 6.}

A canonical game for assessing trust is the investment game of Berg, Dickhaut and McCabe (1995). Two players are anonymously paired, and each is given an initial endowment of money. One of the players (the principal) has to choose a portion of his endowment to send to the other player (the agent), and the principal is told that the experimenter will multiply the investment by three. For example, if the principal invests 1 rupee, then the agent receives 3 rupees. Finally, the agent chooses a non-negative amount of money to return
to the principal. A self-interested agent will always seize the investment, so the unique Nash equilibrium would be that the principal invests nothing. A strictly positive investment is interpreted as an indication that the principal trusts the agent to pay a return.

We implemented a binary choice version of this game in 2007 in the district of Unnao, Uttar Pradesh. Unnao is approximately 40 miles from the district in which we conducted the Stag Hunt experiment. The players in a given pairing were men drawn from different villages within the district. Each player was given an endowment of 50 rupees, which was equivalent to the daily wage for an unskilled worker. The principal had to choose between investing all or none of the 50 rupees. If the principal chose to invest, the agent received an additional 150 rupees so that he had a total of 200. Then the agent chose whether or not to send back 100 rupees. There were four pairings that differed in the caste statuses of the two players. In LL pairs, both players were low-caste, and HH pairs, both players were high-caste. For LH pairs, the principal was low-caste and the agent was high-caste. In HL pairs, the principal was high-caste and the agent was low-caste. Statuses were indicated to the players by the use of names that were known to be exclusive to a particular caste, as we verified in a pre-experiment test in the district. The numbers of pairs by caste status were: 26 (LL), 34 (LH), 30 (HH), and 30 (HL).

The left panel of Figure 7 presents the proportion of principals who invested. There are no significant differences in this proportion across different pairs. In fact, trust as measured by the proportion of principals who invest is higher for HH (80 percent) than for the other types of pairs.

---

17Hoff, Kshetramade and Fehr (2011) report a version of this game with third-party punishment.
We also asked the principals: Do you believe that the agent will send you back 100 rupees if you invest? The right panel of Figure 7 reports the proportion that said yes. There are no significant differences in principals’ beliefs by the caste status of the principal or the agent. Finally, we asked the agents: Do you believe the principal will invest the 50 rupees? In a probit regression (not reported here), all coefficients are close to zero and insignificant, and the results are robust to controls for the agents’ wealth.

In sum, the investment game experiment provides no evidence that high-caste men are less trusting than low-caste men. The results suggest that trust differences do not account for the caste differences in behavior in the Stag Hunt.

6.2 A Difference in Self-efficacy?

Another alternative explanation of the results in the Stag Hunt would be that high- and low-caste cultures have the same mental models for interpreting others’ actions and the same rules of thumb for retaliation, but that low-caste men feel unable to respond to transgressions. The belief in the ability to have an effect on one’s environment has been variously referred to as locus of control or self-efficacy in the psychology literature. If low-caste individuals systematically had a lower sense of self-efficacy, then that would explain the results of the Stag Hunt experiment and the vignette-based survey.

To investigate this hypothesis, we conducted an additional survey in 2013 involving 62 men—30 high-caste and 32 low-caste—to test for caste differences in self-efficacy. We employed a survey design from the psychology literature (Paulhus, 1983): The men were presented with four statements that take positions on different aspects of one’s locus of control, e.g., “I have no trouble making and keeping friends.” The respondents were asked whether or not they felt that the statements described them.

Table 5 reports the statements and the proportions of men who responded yes or no to whether or not the statements described them. The remaining respondents gave equivocal answers or were unable to respond. Not only did the responses not vary significantly across castes, but the point estimates indicate that if anything, low-caste men have greater locus of control than their high-caste counterparts. For example, the second row shows that 84 percent of low-caste men disagreed that they had a hard time finding others to help them, compared to 77 percent of high-caste men. The results suggest that a caste difference in self-efficacy does not explain the caste difference in responses in the punishment survey.
| Does this statement describe you?                                                                 | Yes | No | N  |
|-------------------------------------------------------------------------------------------------|-----|----|----|
|                                                                                                 | L   | H  | L  | H  |    |
| I have no trouble making and keeping friends.                                                   | 0.97| 0.97| 0.00| 0.03| 62 |
| If I need help in carrying off a plan of mine, it’s usually difficult to get others to help.    | 0.06| 0.13| 0.84| 0.77| 62 |
| I often find it hard to get my point of view across to others.                                  | 0.03| 0.17| 0.78| 0.77| 62 |
| In attempting to smooth over a disagreement, I usually make it worse.                           | 0.09| 0.07| 0.84| 0.83| 62 |

Table 5: Responses to the survey on self-efficacy.

7 Conclusion

There is increasing awareness that one should not draw inferences about universals in human behavior from the results of experiments with US college students. This paper provides a new example in behavioral economics in which results do not generalize and in which cultural traits have important economic consequences. It presents results from an experiment in coordination with high-caste and low-caste men in ten villages in northern India. In general, the low-caste men coordinated on the efficient convention, just as experimental results with US college students led us to expect, but the high-caste men did not—either when they were interacting among themselves or with members of low castes, though convergence to efficiency was more likely with members of low castes. Under the efficient convention, the endowment in each period grew by 67 percent (from 6 rupees to 10 rupees), whereas under the inefficient convention, it grew by only 17 percent (from 6 rupees to 7 rupees). The disparity in outcomes may be due to the culture of honor. Because the high castes adhere to the culture of honor much more strongly than the low castes, they may interpret the loser’s payoff and other affronts as challenges to their honor or status. The culture of honor is associated with the mindset, “Cross me and I’ll punish you,” under which accidental miscoordination gives rise to retaliation and more miscoordination and misunderstanding and the unraveling of attempts at collaboration. Almost 90 percent of both high- and low-caste players continued to take the collaborative action, Stag, after reaching the efficient equilibrium in the Stag Hunt. There is thus no evidence that high- or low-caste men seek to “insult” their partners by imposing a loss on them. However, the mental representations that high-caste men create to interpret the loss they incur from miscoordination (the loser’s payoff) appear to lead them to this biased view. Our results are robust to controls for education and wealth, which
suggests that the covariates of caste do not explain the caste gap in coordination. We found no support for two other potential explanations: caste differences in trust and self-efficacy.

The poor ability of high-caste men to coordinate efficiently in the Stag Hunt is consistent with the observed pervasive failures in coordination and institution-building in Uttar Pradesh, which Drèze and Gazdar (1997, p. 101) characterize as “social and political inertia.”

Our theoretical framework gives new micro-foundations for psychological processes that impede economic development and suggests mechanisms to target with interventions. There is evidence from large randomized controlled trials (RCTs) in Chicago and in rural areas of Liberia that the perception of challenges to one’s status, as well as the automatic response of aggressive retaliation, are malleable. In Chicago, Heller et al. (2015) evaluate three RCTs of programs among disadvantaged male youth to reduce automatic retaliation to assertions of authority and to perceived affronts. Such responses may be adaptive in “street life” in disadvantaged areas but are ill-suited to other environments. The programs thus did not discourage participants from retaliating against threatening assertions of authority, but rather encouraged participants to recognize when automatic responses might be maladaptive and to think about what response was called for. For example, when a teacher tells him to sit down and be quiet so class can start, a participant would learn not to “feel like one’s reputation is at stake” (p. 4). An intervention that helped very disadvantaged high school boys in 18 pubic school think deliberately had the effect of reducing their violent-crime arrests in that year by 44 percent (p. 22). A study of another such intervention in a high-risk juvenile detention center found that in the control group, only one-quarter of the young men had not been readmitted to the detention center within 18 months, whereas the treatment increased that rate by two-thirds (p. 27). In post-civil war rural Liberia, Blattman et al. (2012) found that a training program in mediation gave people new conceptual tools for understanding conflict—for instance, seeing the interaction as positive-sum instead of zero-sum—and new norms that improved coordination. A survey of over 10,000 people one year after the intervention found that the training program did not reduce dispute levels, but it reduced the level of unresolved conflicts and increased the proportion of people satisfied with their resolution.

In these interventions, large social changes were brought about by more or less simultaneous frame switches by many individuals, a mechanism of social change highlighted by DiMaggio (1997). Interventions can target and change the mental models through which people interpret the situations they are in, and the responses that their interpretations trigger. Mental models and rules of thumb are understudied in economics—a gap that our study and other recent work tries to address.
References

Bacharach, Michael, “Variable Universe Games,” in Ken Binmore, Alan Kirman, and Piero Tani, eds., *Frontiers of Game Theory*, Cambridge, MA: MIT Press, 1993, pp. 255–275.

Basu, Kaushik, “The republic of beliefs: A new approach to ‘law and economics’,” Policy Research Working Paper No. 7259, World Bank 2015.

Bénabou, Roland, “Ideology,” *Journal of the European Economic Association*, 2008, 6 (2–3), 321–352.

Benaïm, Michel and Morris W Hirsch, “Mixed equilibria and dynamical systems arising from fictitious play in perturbed games,” *Games and Economic Behavior*, 1999, 29 (1), 36–72.

Berg, Joyce, John Dickhaut, and Kevin McCabe, “Trust, reciprocity, and social history,” *Games and economic behavior*, 1995, 10 (1), 122–142.

Bertrand, Marianne, Sendhil Mullainathan, and Eldar Shafir, “A behavioral-economics view of poverty,” *The American Economic Review*, 2004, 94 (2), 419–423.

Binmore, Ken and Larry Samuelson, “An economist’s perspective on the evolution of norms,” *Journal of Institutional and Theoretical Economics (JITE)/Zeitschrift für die gesamte Staatswissenschaft*, 1994, pp. 45–63.

Blattman, Christopher, Alexandra Hartman, and Robert Blair, “Building institutions at the micro-level: Results from a field experiment in property dispute and conflict resolution,” *Available at SSRN 2158966*, 2012.

Charness, Gary and Matthew Rabin, “Understanding social preferences with simple tests,” *The Quarterly Journal of Economics*, 2002, 117 (3), 817–869.

D’Andrade, Roy, *The development of cognitive anthropology*, Cambridge: Cambridge University Press, 1995.

Derné, Steve, “Cultural conceptions of human motivation and their significance for culture theory,” in Diana Crane, ed., *The sociology of culture*, Oxford: Blackwell, 1994, pp. 267–287.

DiMaggio, Paul, “Culture and cognition,” *Annual review of sociology*, 1997, 23, 263–287.

_ and Hazel Rose Markus, “Culture and Social Psychology: Converging Perspectives,” *Social Psychology Quarterly*, 2010, 73 (4), 347–352.

Dixit, Avinash, “An assurance model of ‘anti-corruption’,” Working Paper, Princeton University 2016.
Drèze, Jean and Haris Gazdar, “Uttar Pradesh: The burden of inertia,” in Jean Drèze and Amartya Sen, eds., Indian development: Selected regional perspectives, Oxford: Oxford University Press, 1997, pp. 33–128.

_ and Naresh Sharma, “Palanpur: population, society, economy,” in Peter Lanjouw and Nicholas Stern, eds., Economic Development in Palanpur over Five Decades, Oxford: Oxford University Press, 1998, pp. 3–113.

Ellickson, Robert C, Order without law: How neighbors settle disputes, Harvard University Press, 1991.

Fehr, Ernst and Klaus M Schmidt, “A theory of fairness, competition, and cooperation,” The Quarterly Journal of Economics, 1999, 114 (3), 817–868.

Fiske, Alan Page, Shinobu Kitayama, Hazel Rose Markus, and Richard E Nisbett, “The cultural matrix of social psychology,” in Daniel T Gilbert, Susan T Fiske, and Gardner Lindzey, eds., The handbook of social psychology, New York: McGraw-Hill, 1998, pp. 915–981.

Friedrich, Paul, “Sanity and the Myth of Honor,” Ethos, 1977, 5 (3), 281–305.

Fudenberg, Drew and David K Levine, The theory of learning in games, Vol. 2, Cambridge: MIT press, 1998.

Gregg, Gary S, The Middle East: a cultural psychology, Oxford University Press, 2005.

Greif, Avner, Institutions and the path to the modern economy: Lessons from medieval trade, Cambridge: Cambridge University Press, 2006.

Grosjean, Pauline, “A history of violence: The culture of honor and homicide in the US south,” Journal of the European Economic Association, 2014, 12 (5), 1285–1316.

Haidt, Jonathan, Silvia Helena Koller, and Maria G Dias. “Affect, culture, and morality, or is it wrong to eat your dog?,” Journal of personality and social psychology, 1993, 65 (4), 613.

Heller, Sara B, Anuj K Shah, Jonathan Guryan, Jens Ludwig, Sendhil Mullainathan, and Harold A Pollack, “Thinking, Fast and Slow? Some Field Experiments to Reduce Crime and Dropout in Chicago,” WP21178, National Bureau of Economic Research 2015.

Herrmann, Benedikt, Christian Thöni, and Simon Gächter, “Antisocial punishment across societies,” Science, 2008, 319 (5868), 1362–1367.

Hoff, Karla and Joseph E Stiglitz, “Equilibrium Fictions: A Cognitive Approach to Societal Rigidity,” The American Economic Review, 2010, 100 (2), 141–146.

_ and _ , “Striving for balance in economics: Towards a theory of the social determination of behavior,” Journal of Economic Behavior and Organization, 2016, 126, 25–57.
Mayuresh Kshetramade, and Ernst Fehr, “Caste and Punishment: the Legacy of Caste Culture in Norm Enforcement,” *The Economic Journal*, 2011, 121 (556), F449–F475.

Huyck, John B Van, Raymond C Battalio, and Richard O Beil, “Tacit coordination games, strategic uncertainty, and coordination failure,” *The American Economic Review*, 1990, 80 (1), 234–248.

Jackson, Matthew O and Yiqing Xing, “Culture-dependent strategies in coordination games,” *Proceedings of the National Academy of Sciences*, 2014, 111 (Supplement 3), 10896–10896.

Kahneman, Daniel and Amos Tversky, “Choices, values, and frames,” *American Psychologist*, 1984, 39 (4), 341.

Khare, R. S., *The untouchable as himself: Ideology, identity, and pragmatism among the Lucknow Chamars*, Cambridge: Cambridge University Press, 1984.

Knez, Marc and Colin Camerer, “Creating expectational assets in the laboratory: Coordination in ‘weakest-link’ games,” *Strategic Management Journal*, 1994, 15 (S1), 101–119.

Kushner, Harold J and George Yin, *Stochastic approximation and recursive algorithms and applications*, Vol. 35, New York: Springer-Verlag, 2003.

Lewis, David, *Convention: A philosophical study*, Harvard University Press, 1969.

Mandelbaum, David G, *Women’s seclusion and men’s honor: Sex roles in North India, Bangladesh, and Pakistan*, Tucson: University of Arizona Press, 1993.

Nisbett, Richard E and Dov Cohen, *Culture of honor: The psychology of violence in the South*, Colorado: Westview Press, 1996.

Nunn, Nathan, “Culture and the historical process,” *Economic History of Developing Regions*, 2012, 27 (sup1), 108–126.

Paulhus, Delroy, “Sphere-specific measures of perceived control,” *Journal of Personality and Social Psychology*, 1983, 44 (6), 1253.

Pettigrew, Joyce, “The Jats of Punjab,” in Dipankar Gupta, ed., *Social stratification*, New Delhi: Oxford University Press, 1991, pp. 163–175.

Pitt-Rivers, Julian, “Honour and social status,” in J.G. Peristiany, ed., *Honour and shame: The values of Mediterranean society*, Chicago: University of Chicago Press, 1966, pp. 19–77.

Schneider, Jane, “Of vigilance and virgins: Honor, shame and access to resources in Mediterranean societies,” *Ethnology*, 1971, 10 (1), 1–24.

Skyrms, Brian, *The stag hunt and the evolution of social structure*, Cambridge: Cambridge University Press, 2003.
Swidler, Ann, “Culture in action: Symbols and strategies,” *American Sociological Review*, 1986, 51 (2), 273–286.

Weber, Roberto A, “Managing Growth to Achieve Efficient Coordination in Large Groups,” *American Economic Review*, 2006, 96 (1), 114–126.

World Bank, *World Development Report 2015: Mind, Society, and Behavior*, World Bank, 2015.
Formal Analysis of Learning Model

This Appendix contains the formal analysis of the model of Section 5.

First we make some preliminary observations about \( b \). Using the utilities in Figure 1, we can calculate \( u_i(\text{Stag}|p, I) = 10(1 - \sigma_i)p + (3 - \sigma_i)(1 - p) \) and \( u_i(\text{Hare}|p, I) = (7 - 3\sigma_i)p + 7(1 - \sigma_i)(1 - p) \). Hence, \( u_i(\text{Stag}|p, I) - u_i(\text{Hare}|p, I) = 7(1 - \sigma_i)p - 4 \) and so \( b_i(p|N) = (1 + \exp(4 - 7p))^{-1} \) and \( b_i(p|I) = (1 + \exp(4 - 7(1 - \sigma_i)p))^{-1} \). We note for future reference that \( b_i(p|I) \) is decreasing in \( \sigma_i \), so that \( \Delta_i(p) = b_i(p|N) - b_i(p|I) \) is increasing in \( \sigma_i \).

We begin by solving explicitly for \( \Gamma(s_i, s_j|p, i) \) and \( b(p_i, p_j) \) for a fixed \( \sigma_1, \sigma_2 \). In the following expressions, we will economize on notation by writing \( b_i \) for \( b_i(p_i|N) \) and \( \Delta_i \) for \( \Delta_i(p_i) \). \( \Gamma \) and \( b_i \) are given by

\[
\begin{align*}
\Gamma(I, N|p_i, p_j) &= \frac{b_i(1 - b_j + \Delta_j(1 - b_i))}{(1 + \Delta_i)(1 + \Delta_j) - b_i\Delta_j(1 + \Delta_i) - b_j\Delta_i(1 + \Delta_j)} \quad (6a) \\
b_i(p_i, p_j) &= \frac{b_i(1 + \Delta_j(1 - b_i) - b_j\Delta_i(1 + \Delta_j))}{(1 + \Delta_i)(1 + \Delta_j) - b_i\Delta_j(1 + \Delta_i) - b_j\Delta_i(1 + \Delta_j)} \quad (6b)
\end{align*}
\]

\( \Gamma \) can be derived from the system of equations

\[
\begin{align*}
\Gamma(I, N|p_i, p_j) &= \Gamma(I, N|p_i, p_j)b_i(p_i|I)(1 - b_j(p_j)) \\
&\quad + \Gamma(N, I|p_i, p_j)b_i(p_i|N)(1 - b_j(p_j|I)) \\
&\quad + (1 - \Gamma(I, N|p_i, p_j) - \Gamma(N, I|p_i, p_j))b_i(p_i|N)(1 - b_j(p_j|N))
\end{align*}
\]

and the analogous equation flipping the identities \( i \) and \( j \). The details are left to the reader.

From these formulae, the expression for \( b_i(p_i, p_j) \) can be derived from

\[
b_i(p_i, p_j) = b_i(p_i|N) - \Delta_i(p_i)\Gamma(I, N|p_i, p_j)
\]

Using this expression, we can derive comparative statics for \( b_i \) in \( \sigma_i \) and \( \sigma_j \). Note that \( b_i(p_i, p_j) \) can be rewritten

\[
b_i(p_i, p_j) = \frac{1 + \Delta_j(1 - b_i - b_j\Delta_i)}{1 + \Delta_j(1 - b_i - b_j\Delta_i) + \Delta_i(1 - b_j + \Delta_j(1 - b_i))}.
\]

An increase in \( \sigma_i \) or \( \sigma_j \) only affects \( b_i(p_i, p_j) \) through increases in \( \Delta_i \) and \( \Delta_j \), respectively. Thus, we can look for comparative statics of \( \frac{X}{X+Y} \) as either \( \Delta_i \) or \( \Delta_j \) increase. Note that \( \frac{X}{X+Y} \) decreases only if \( \frac{Y}{X} \) increases. It is easy to see that \( X \) is decreasing in \( \Delta_i \) and \( Y \) is
increasing in \( \Delta_i \) (since \( 1 - b_i, 1 - b_j, \) and \( \Delta_j \) are all positive). Hence, an increase in \( \sigma_i \) must lead to a decrease in \( b_i(p_i, p_j) \). For \( \Delta_j \), the argument is only slightly more involved. To show that \( b_i(p_i, p_j) \) is decreasing in \( \sigma_j \), it is sufficient to show that \( \frac{d}{d\Delta_j} \) is increasing in \( \Delta_j \), which is the case if \( X \frac{dV}{d\Delta_j} - Y \frac{dX}{d\Delta_j} > 0 \). But this expression reduces to \( Y \Delta_j b_j + \Delta_i (1 - b_i)b_j (1 - \Delta_i \Delta_j) \), which must be positive since \( \Delta_i \Delta_j < 1 \). Thus, \( b_i(p_i, p_j) \) is decreasing in both \( \sigma_i \) and \( \sigma_j \).

An important property for characterizing the model is that the functions \( p_i - b_j(p_j, p_i) \) and \( b_i(p_i, p_j) \) should both be increasing in \( p_i \). This property does not hold for all possible specifications of the players’ utilities, but it is true for the particular payoffs used in our experiment. An analytical proof of this fact would be both lengthy and unedifying, so we have simply verified numerically that it holds in all of the values of \( \sigma_i \) utilized in our examples, and we suspect it holds more generally. For the rest of the analysis, this property is treated as an assumption.

Kushner and Yin (2003, Chapter 8 Theorem 4.3) shows that the forecasts converge to a trajectory of the dynamical system (5). A consequence of the fact that \( p_i - b_j(p_j, p_i) \) is increasing is that \( \frac{\partial}{\partial p_i} \left( b_i(p_i, p_j) \right) < 0 \). Hence, the divergence of the vector field for \( (p^1, p^2) \) is negative, and the vector field is area decreasing. Therefore, no non-empty open sets are invariant under the mean dynamical system, and beliefs must converge to a stable steady state of (5) (cf. Benaïm and Hirsch, 1999; Fudenberg and Levine, 1998).

In addition, the fact that \( p_i - b_j(p_j, p_i) \) is strictly increasing means that there exists a unique solution to the functional equation \( p_i^*(p_j) = b_j(p_j, p_i^*(p_j)) \). For \( b_j(p_j, p_i) \in (0, 1) \) and hence \( 0 - b_j(p_j, 0) < 0 \) and \( 1 - b_j(p_j, 1) > 0 \), and the expression is strictly increasing and continuous, so by the intermediate value theorem, a solution to \( p_i - b_j(p_j, p_i) = 0 \) must uniquely exist. For \( p_i < p_i^*(p_j), p_i < b_j(p_j, p_i), \) and for \( p_i > p_i^*(p_j), p_i > b_j(p_j, p_i) \).

Moreover, because \( b_j(p_j, p_i) \) is strictly increasing in \( p_j \), it must be that \( p_i^* \) is strictly increasing in \( p_j \) as well and has a strictly increasing inverse \( p_i^{*-1} \). For if \( p'_j > p_j \), then \( 0 = p_i^*(p_j) - b_j(p_j, p_i^*(p_j)) > p_i^*(p_j) - b_j(p'_j, p_i^*(p_j)) \), and since \( p_i - b_j(p_j, p_i) \) is strictly increasing, we must have that \( p_i^*(p'_j) > p_i^*(p_j) \). By a similar argument, since \( b_i(p_i, p_j) \) is decreasing in \( \sigma_i \) and \( \sigma_j \), \( p_i^* \) must be decreasing in \( \sigma_i \) and \( \sigma_j \) as well.

Finally, we need the following technical result to characterize the equilibria:

**Lemma 1.** Let \( h^1, h^2, l^1, l^2 : [0, 1] \rightarrow (0, 1) \) be continuous and strictly increasing functions, with \( h^1 \geq l^1 \) and \( h^2 > l^2 \). Let \( \bar{x}_h \) denote the largest solution to \( h^1(x) = h^2^{-1}(x) \) and let \( \underline{x}_h \) be the smallest solution to \( l^1(x) = l^2^{-1}(x) \). Then for any solutions \( x_h \) and \( x_l \) to \( h^1(x) = h^2^{-1}(x) \) and \( l^1(x) = l^2^{-1}(x) \), respectively, we must have \( x_l < \bar{x}_h \) and \( \underline{x}_h < x_h \).

**Proof.** Let \( \bar{x} = (h^2)^{-1}(1) \). Observe that since \( h^2 > l^2, (l^2)^{-1} > (h^2)^{-1}, \) as \( h^2(x) = l^2(x') = y \) implies that \( x < x' \).
For $x > \bar{x}_h$, it must be that $h^1(x) < (h^2)^{-1}(x)$. For $h^1(\bar{x}) < (h^2)^{-1}(\bar{x}) = 1$, so if $h^1(x) > (h^2)^{-1}(x)$, then by continuity there is an $x' \in (x, \bar{x})$ such that $h^1(x') = (h^2)^{-1}(x')$, and $x' > \bar{x}_h$, which contradicts that $\bar{x}_h$ is the largest solution. But if $\bar{x}_h \leq x_l$, then $(h^2)^{-1}(x_l) < (l^2)^{-1}(x_l) = l^1(x_l) < h^1(x_l)$.

The proof that $\underline{x} < x_h$ is essentially the same. \[\square\]

This lemma implies that the set of learning equilibria are strictly decreasing in the weak set order: the functions $p^*_1$ and $p^*_2$ are strictly decreasing in $(\sigma_1, \sigma_2)$, so that both $\bar{p}(\sigma_1, \sigma_2)$ and $\underline{p}(\sigma_1, \sigma_2)$ are decreasing as well.

We can verify numerically that $p^*_1(\cdot|1,1)$ and $p^*_2^{-1}(\cdot|1,1)$ have a single point of intersection. By continuity, there will exist an open set around $(1,1)$, denoted $\Sigma \in [0,1]^2$, such that if $(\sigma_1, \sigma_2) \in \Sigma$, there is a unique point of intersection $\bar{p}(\sigma_1, \sigma_2)$ that is lower than the lowest point of intersection $\underline{p}(0,0)$.
Online Appendix

All vignettes except V4 use caste neutral names. In the control vignette 0, the offense is unambiguous. The law is violated. The law prescribes the appropriate response. The vignettes 1-3 have some ambiguity, and thus there is greater scope for cultural attitudes towards punishment to operate. It is a difference in caste culture that we seek to evaluate. V4 is about inter-caste marriage, a central issue in the culture of honor in India. No mention of caste is made except in V4.

V0: (Control) A thief named Dinesh robs Mahesh’s house at night. Mahesh finds out, beats up the thief, and reports him to the police. The thief is locked up in the police lock-up for a few days.

V1: Dinesh digs a canal through the field of Mahesh to have water reach his field. Dinesh does not obtain permission from Mahesh before doing this. He digs over night and Mahesh finds out the following morning, only after the canal has been dug. After this incident, Mahesh met Dinesh in public, had an argument, and badly beat up Dinesh.

V2: Dinesh lets his cattle graze on Mahesh’s fields at night. Mahesh’s crop is ready for harvest. Dinesh does not inform Mahesh, nor does he obtain permission from Mahesh before letting his cattle graze in Mahesh’s fields. Mahesh finds out only after the crop has been eaten by Dinesh’s animals. After this incident, Mahesh met Dinesh in public, had an argument, and badly beat up Dinesh.

V3: Mahesh was going back home at night for a feast in the village. On the way, he met Dinesh sitting with his friends. Dinesh called Mahesh names and beat him up. After this incident, Mahesh met Dinesh in public, had an argument, and badly beat up Dinesh.

V4a: Dinesh Thakur’s daughter marries Mahesh Bania. After this incident, Dinesh met Mahesh in public, had an argument, and badly beat up Mahesh.

V4b: Dinesh Pasi’s daughter marries Mahesh Chamar. After this incident, Dinesh met Mahesh in public, had an argument, and badly beat up Mahesh.

Note: V4a is for high-caste subjects and V4b is for low-caste subjects. Thakur and Bania are high castes, with Bania being a lower rank. Similarly, Pasi and Chamar are low castes, with Chamar being a lower rank.

Questions for the respondents:
(1) Was Mahesh’s action justified?

(2) If you were in Mahesh’s place, what would you have done?