When does reputation lie? Dynamic feedbacks between costly signals, social capital, and social prominence

Marion Dumas
Grantham Research Institute, London School of Economics & Political Science

Jessica L. Barker
Aarhus University Interacting Minds Centre, Alaska Dept of Health & Social Services

Eleanor A. Power
London School of Economics & Political Science

July 19, 2021

Abstract

Performing a dramatic act of religious devotion, creating an art exhibit, or releasing a new product are all examples of public acts that signal quality and contribute to building a reputation. Signalling theory predicts that these public displays can reliably reveal quality. However, data from ethnographic work in South India suggests that more prominent individuals gain more from reputation-building religious acts than more marginalised individuals. To understand this phenomenon, we extend signalling theory to include variation in people's social prominence or social capital, first with an analytical model and then with an agent-based model. We consider two ways in which social prominence/capital may alter signalling: (1) it impacts observers' priors, and (2) it alters the signallers' payoffs. These two mechanisms can result in both a "reputational shield," where low quality individuals are able to "pass" as high quality thanks to their greater social prominence/capital, and a "reputational poverty trap," where high quality individuals are unable to improve their standing due to a lack of social prominence/capital. These findings bridge the signalling theory tradition prominent in behavioural ecology, anthropology, and economics with the work on status hierarchies in sociology, and shed light on the complex ways in which individuals make inferences about others.

Keywords: costly signalling, reputation, Matthew Effect, social capital, poverty trap

1 Introduction

Why do scientific publications published according to the same peer-review standard receive different levels of attention depending on whether their authors have been highly cited in the past (Hirsch, 2007)? Why can one winery sell a bottle of wine for $2000 while another winery can only sell a comparable bottle for $2 (Benjamin and Podolny, 1999)? Why is the boost in reputation that a villager enjoys after taking part in a ritual at a religious festival higher if the villager is well connected than if the villager is more isolated in the village's social network

(Power, 2015)? These disparate examples share a common structure: individuals' perceptions of another's quality depend on his or her social prominence; that is, how well-regarded he or she is. (We use the term "social prominence" as an umbrella category to refer to what has elsewhere been called status, prestige, or dominance).

Within sociology, there is a large body of literature, generally overlooked by evolutionary scientists, that focuses on how status (the most commonly used term for "social prominence" in this field) and quality can become disconnected. A common example of this is referred to as the "Matthew Effect," the idea that the "rich get richer," as more prominent individuals receive more recognition for their work than do less prominent individuals, regardless of underlying quality (Merton, 1968). Evidence for the cumulative advantage of status has been found in arenas as distinct as academia (e.g., Newman, 2009; Simcoe and Waguespack, 2010; Petersen et al., 2011), the wine industry (Benjamin and Podolny, 1999), and music (Salganik et al., 2006). Sociologists have documented many examples of inequality and "status dispersion" that seem to indicate a gap between underlying quality or merit and the reputational rewards people ultimately receive (e.g., Menger, 1999; Oakley and O'Brien, 2016). Experimental work (Ridgeway and Erickson, 2000; Salganik et al., 2006; Muchnik et al., 2013; van de Rijt et al., 2014; Correll et al., 2017; Hackel and Zaki, 2018) has shown how such patterns can emerge particularly in situations where individuals draw on the beliefs of others. These empirical findings, and some formal models (e.g., Podolny, 1993; Lynn et al., 2009; Manzo and Baldassarri, 2015), suggest that the use of social prominence when attempting to evaluate quality may lead to self-reinforcing dynamics that ultimately decouple the two.

Within evolutionary anthropology, there is a long-held recognition of the many benefits of social prominence (e.g., Irons, 1979; Smith, 2004; von Rueden et al., 2011; Majolo et al., 2012; von Rueden et al., 2014). There is also a growing attention to and evidence for the intergenerational transfer of wealth of all forms, and how it confers sizeable advantages to those born to parents with higher embodied, relational, and material wealth (Borgerhoff Mulder et al., 2009). In other species, such as Japanese macaques (Chapais, 1988) and spotted hyaenas (Engh et al., 2000), there is evidence that dominance rank can sometimes be inherited, leaving rank disconnected from strength. Similarly, within the behavioural ecological work on dominance hierarchies there is increasing recognition of the critical role that social dynamics have on ranks otherwise expected to follow from differences in intrinsic attributes (Chase et al., 2002). The intergenerational transfer of social prominence and the benefits that accrue from it suggest again that there may be cumulative advantages to social prominence.

Despite this empirical evidence, many of the models used within the evolutionary sciences posit a straightforward relationship between the social prominence of an individual and the "quality" of that individual. Depending on the context, "quality" may mean attributes such as fitness, strength, cooperativeness, knowledge, or skill. For example, models of cultural transmission suggest that prestigious individuals receive deference because of the skills and knowledge that they possess (Henrich and Gil-White, 2001; Henrich et al., 2015). Models of indirect reciprocity use the direct history of individuals' actions as a proxy for reputation (e.g., Ohtsuki and Iwasa, 2006; Roberts et al., 2021; Santos et al., 2021), and suggest that gossip about an individual can accurately convey that reputation (e.g., Sommerfeld et al., 2008). Economic and evolutionary signalling models provide a framework within which quality may be accurately assessed through the relative costliness of signals. In the canonical models (Spence, 1973; Grafen, 1990a;b), individuals' assessments of others are tightly linked to quality, although an accumulation of theory building on this early work has shown that the relationship between signal cost, signaller quality, and receivers' perceptions is more complex (e.g., Huttegger et al., 2014; Zollman et al., 2012; Wagner, 2013; Lachmann and Bergstrom, 1998). This theoretical literature, however, has generally not explicitly taken into account social prominence, which the empirical evidence suggests is also implicated in this more muddled relationship between quality and reputation.

Here, then, we engage with a set of tightly linked concepts, all related to different aspects of social evaluation and social connection (cf. Power and Ready, 2018). So far, we have introduced social prominence, which generally has to do with relative standing, and is often marked by acts of deference. To this, we now add the related concept of social capital, which has to do with social connections and the resources they provide, marked not by acts of deference, but by acts of interpersonal support (Bourdieu, 1986; Lin, 2001). Finally, reputation refers to the beliefs that others have about an individual's qualities, based on the assessment of their (observed or reported) actions. Building on the evidence outlined above, we expect that social prominence and social capital may both be drawn into the process of reputation formation and assessment. Exploring the dynamic interplay between these may help to explain the conditions under which we expect them to align or *mis*align.

To do so, we extend the canonical costly signalling model (Spence, 1973; Grafen, 1990a) to include the signaller's social prominence and social capital. This model predicts that by engaging in costly signals, individuals can reveal their quality. We seek to analyse how social prominence and social capital might affect the reputational gains individuals get from engaging in costly signals. We consider two mechanisms by which social prominence or social capital can influence signalling: 1) they impact the observer's evaluation of the signaller, or 2) they directly alter the payoffs of signalling.

In the first case ("altered prior"), we are suggesting that social prominence or social capital may be used as an indicator of quality by observers; if a signalling act provides an opportunity for observers to note the attention, deference, or support received by an individual, then observers may use this social information, alongside the costly signal itself, to update their assessment of the signaller's quality (i.e., the signaller's reputation). The idea that people often interpret social prominence or social capital as an indication of quality is well accepted in sociology (e.g., Podolny, 1993; Burt, 2008; Manzo and Baldassarri, 2015), as well as in evolutionary anthropology (e.g., Henrich and Gil-White, 2001), and social psychology (Cialdini and Trost, 1998).[1]

In the second case ("altered payoff"), we are suggesting that the attention or support that a signaller has may enhance the visibility of a signal or otherwise facilitate its enactment, meaning that the net benefit to more prominent signallers may be higher. Social capital is fundamentally seen as having productive potential (Bourdieu, 1986; Coleman, 1988), and sociologists have outlined the possibility that network effects may contribute to patterns of cumulative advantage (DiMaggio and Garip, 2012; DiPrete and Eirich, 2006), so we should have a strong expectation of this effect[2].

[1] One way to understand this is that social prominence/capital is being used as a cue of quality. We avoid using the term "cue" here, however, because it has a very particular and narrow meaning in the behavioural ecology literature (Maynard Smith and Harper, 1995).

[2] We note that although the behavioural ecological literature has considered differential benefits, from early costly signalling theory to explain begging by chicks (e.g., Godfray, 1991) to more recent models (Whitmeyer, 2021), they consider the *opposite* scenario to the one on which we focus here: that is, where the hungrier chicks beg more and receive more food, unlike the "rich get richer" dynamic exemplified by the case study below.

While we think that both social prominence or social capital could operate with either mechanism, the altered prior mechanism may be more readily associated with social prominence, and the altered payoff mechanism with social capital. We consider these two mechanisms both separately and in combination, first in an analytical model and then in a dynamic agent-based model in which signalling behaviour, social prominence/capital and social interactions co-evolve. Our analytical model shows that if social prominence/capital is used as a prior for quality, then individuals with higher prominence/capital have a greater reputational gain after signalling than individuals with lower social prominence/capital. The agent-based model shows that this mechanism leads to a “reputational shield” (low quality individuals able to maintain good standing thanks to initially high social prominence/capital). The analytical model also illustrates how allowing social prominence/capital to alter payoffs introduces the possibility that high quality individuals may not be able to signal due to their low social prominence/capital, a possibility which does not exist in the classical costly signalling model. The agent-based model model shows that this mechanism leads to a “reputational poverty trap” (high quality individuals who are unable to improve their standing). The “reputational shield” and “reputational poverty trap” can coexist if both mechanisms are present. By re-examining the sociological models of cumulative advantage in light of signalling theory, we seek to unite the signalling-based approach from behavioural ecology and economics with the sociological theory of status hierarchies.

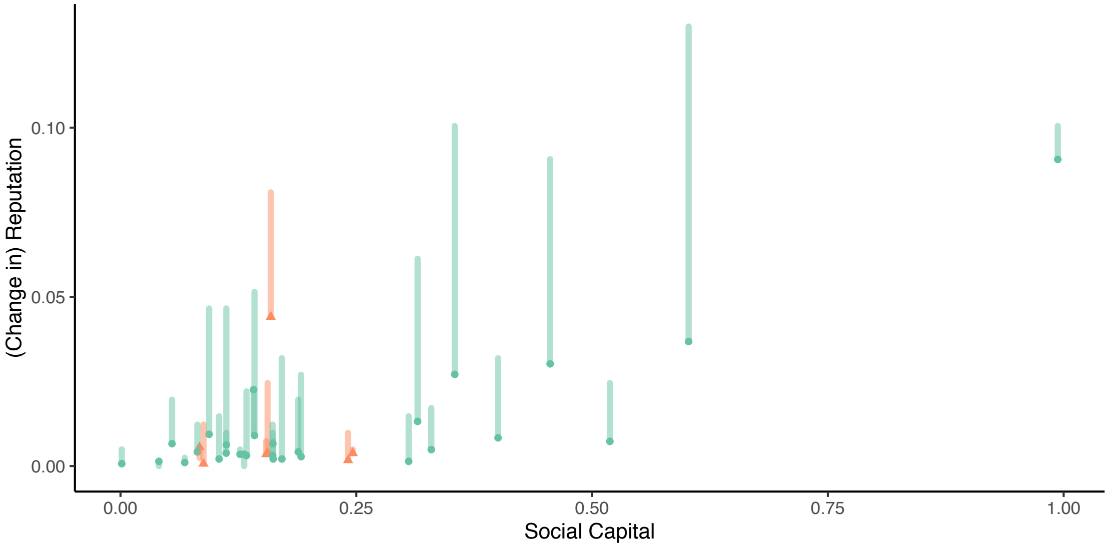


Figure 1: Reputation (as measured by the proportion of possible reputational nominations received) from before (the origin of each point) to after (the terminus of the line) the annual festival, for villagers who perform acts of vow-fulfillment, as a function of social capital (here measured by in-degree centrality in the social support network and normalised, to align with S_i). Orange denotes Dalit (Scheduled Caste) individuals, and green all others.

1.1 Case study: religious signalling in South India

Throughout this paper, we draw upon the example of religious signalling in a village in South India (Power, 2017a;b). At the annual festival for the village goddess, Hindu residents fulfil vows made to the goddess in gratitude for her divine assistance. These vows can range from breaking coconuts to sacrificing animals to walking across a bed of hot coals to piercing one's body with 101 spears. These acts are broadly seen as being revealing of the devotion (*bhakti*) and character (*kuṇam*) of the vow-takers, as they require not just physical endurance but also mental resolve. Those with devotion and strong commitment are understood to bear the burden of these acts more readily than others. While vow fulfilments are generally carried off without a hitch, they can "fail," as when a coconut fails to break when thrown or when a person trips and falls while firewalking. These failures are often interpreted as divine punishment for some fault, meaning that these acts are seen as risky, especially by those unsure of the depth of their own devotion or worthiness. Vow-takers' family and friends often accompany them as they fulfil their vows, helping them to support the weight of the spears, giving them water, or showing them respect by placing a garland on their shoulders. These acts are therefore also opportunities for observing vow-takers' social capital (as seen in the number of accompanying supporters and the density of the pile of garlands around the devotee's neck (Mines, 2005: 162-167)) and social prominence (as seen in the ordered distribution of honours (*mariyātai*) that rank devotees by their status (Appadurai and Breckenridge, 1976)).

Drawing on signalling theory, such acts of vow fulfilment can be seen as credible demonstrations of commitment to the religious community and its moral tenets (e.g., Sosis and Alcorta, 2003; Henrich, 2009; Przepiorka and Diekmann, 2021). Such information may be useful in determining who to form supportive relationships with. Consistent with this, residents who invest more in the religious life of the village are seen as more devout and more prosocial (Power, 2017a), and are more likely to be named as providing others with support (Power, 2017b). These individuals thus appear to benefit through both improved reputations and supportive relationships with others. However, while these social benefits do exist, the size of the benefit varies between individuals (Power, 2015; Xygalatas et al., 2021). Specifically, among the vow-takers who perform particularly dramatic and demanding acts (such as firewalking) during the annual festival, those who already have higher social capital and greater preexisting reputational standing receive relatively larger reputational boosts than those who are less well positioned (see Figure 1), indicating that there may exist increasing reputational returns to social capital.

Throughout this paper, we draw on the example of this signalling system to motivate and interpret the models developed here. We note, however, that we expect the phenomena we are modelling to be quite general: costly signals modified by social prominence/capital have the potential to be quite common in humans, and likely also happen in other group-living species that draw on social information.

2 Analytical model: costly signalling game with social prominence/capital

Before introducing the role of social prominence/capital, we present the basic elements of the model. We posit a game with N individuals who differ in some underlying quality $q \in \{0, 1\}$, 0 for low quality and 1 for high quality. Initially, everyone holds the same prior belief about i's quality, denoted π_{1i} with $\pi_{1i} = 0.5$.

During a public event, each individual simultaneously has the opportunity to engage in a costly signal. Denoting a_i an individual's chosen action, we have $a_i \in (r, \neg r)$, where r is the action to signal and $\neg r$ the action not to signal. The cost of signalling depends on the individual's quality, where c_1 is the cost for an individual of quality $q = 1$ and c_0 is the cost for an individual of quality $q = 0$, with $c_0 > c_1$. For each player i who decides to signal, a move of nature decides the outcome $o_i \in \{s, \neg s\}$ of their signal. Namely, the signal can succeed ($o_i = s$) or fail ($o_i = \neg s$). The probability of success depends on quality, where θ_1 and θ_0 are the probabilities of success for $q = 1$ and $q = 0$ individuals respectively, with $\theta_1 > \theta_0$. The signal and its outcome are public and are observed costlessly. Thus, all players observe all the other players' decision to signal. Having observed this, they make the same inference $\hat{q}_i|a_i, o_i$ about i's quality, conditional on the decision to signal and the signal's outcome.

Having separate term for success/failure is not a usual feature of signalling models, but it is implicit in many signals, as when a vow-taker trips while firewalking, a big game hunter returns empty-handed, or a gazelle staggers instead of stotting properly. Some models have considered errors in signal fidelity (Lachmann and Bergstrom, 1998), where the signal received is not exactly the same as the signal sent, but in this scenario the probability of success is not linked to quality. Others have more explicitly considered success/failure (Huttegger et al., 2015), but differ in that receivers are unable to distinguish between a failed signal and no signal.

The solution concept for this static signalling game is the Bayesian Nash Equilibrium. We allow for mixed strategies and denote $P(r|q = 1)$ and $P(r|q = 0)$ the probabilities that high and low quality individuals respectively engage in the costly signal. Applying Bayes' formula, the inferences $\hat{q}_i$ made by all players $j \neq i$ about i's quality after the public event are:

$$\text{Individual } i \text{ signals and succeeds: } \hat{q}_i|r, s = \frac{\theta_1 P(r|q = 1)\pi_{1i}}{\theta_1 P(r|q = 1)\pi_{1i} + \theta_0 P(r|q = 0)(1 - \pi_{1i})} \tag{1}$$

$$\text{Individual } i \text{ signals and fails: } \hat{q}_i|r, \neg s = \frac{(1 - \theta_1)P(r|q = 1)\pi_{1i}}{(1 - \theta_1)P(r|q = 1)\pi_{1i} + (1 - \theta_0)P(r|q = 0)(1 - \pi_{1i})} \tag{2}$$

$$\text{Individual } i \text{ does not signal: } \hat{q}_i|\neg r = \frac{(1 - P(r|q = 1))\pi_{1i}}{(1 - P(r|q = 1))\pi_{1i} + (1 - P(r|q = 0))(1 - \pi_{1i})} \tag{3}$$

The payoff function is given by $\Pi(q_i, \hat{q}_i, a_i)$. This function satisfies the following properties: (1) signalling is less costly for quality 1 than for quality 0 , which is expressed as $\Pi(1, \hat{q}_i, r) - \Pi(1, \hat{q}_i, \neg r) > \Pi(0, \hat{q}_i, r) - \Pi(0, \hat{q}_i, \neg r)$, and (2) it is beneficial to be perceived to be of high quality, which is expressed as $\frac{d\Pi(q_i,\hat{q}_i,a)}{d\hat{q}_i} > 0$.

We have defined the basic elements of a static game of incomplete information, which we now alter to introduce the role of social prominence/capital via two different mechanisms. For each, we will then solve for the signalling probability for high and low quality individuals (i.e., the Bayesian Nash Equilibrium strategy profiles), as a function of their social prominence/capital.

Mechanism 1: altered prior

Our first intervention is to allow the prior π_{1i} to depend on S_i, our term for either social prominence or social capital. S_i is benchmarked against the individuals with the highest and lowest prominence/capital, so that it ranges from 0 to 1, with the highest value assigned to the individual with the highest prominence/capital. An individual's social prominence/capital is revealed when they signal (regardless of the success or failure of that signal). Using this social information, observers change their prior about the quality of that individual. In other words, multiple streams of information are collected at the event: (1) observation of the signal, (2) its success or failure, and (3) information about the social prominence/capital for those who

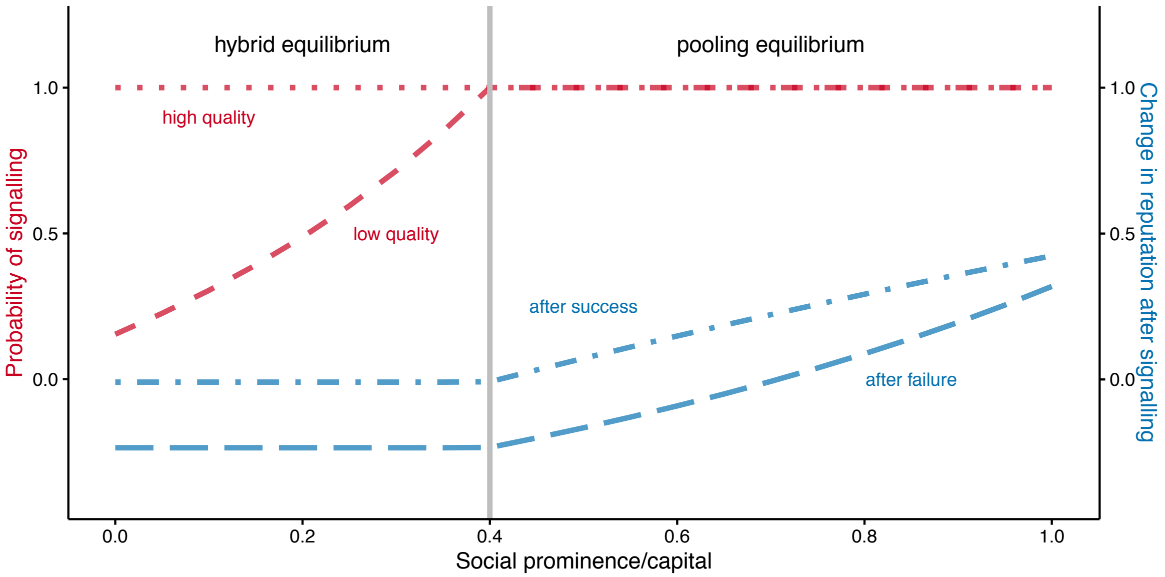


Figure 2: Altered prior mechanism. Red lines indicate the probability of signalling for each type of player (high or low quality). Blue lines show the changes in reputation after signalling, depending on whether the signal failed or succeeded. The parameters used are $c_1 = .2$,$c_0 = .4$,$\theta_1 = .8$,$\theta_0 = .6$,$\Pi = \hat{q}_i - c_i$ and $f(S_i) = .1 + .8S_i$

chose to signal. Mathematically, this means that the prior is a function of S_i, increasing as S_i increases (that is, $\pi_{1i} = f(S_i)$, with $f' > 0$). Importantly, S_i is only observed if the individual signals. Thus, it replaces π_{1i} in Equations 1 and 2 above.

The strategy profiles that form a Bayesian Nash Equilibrium therefore depend not only on quality (which affects the probability of success and the cost), but also on S_i. Thus, in SI 1.2, we solve for the Bayesian Nash Equilibrium strategy profiles $(P^*(r|q = 1, S), P^*(r|q = 0, S))$ with $S \in [0, 1]$, and demonstrate the different signalling regimes that can arise.

Figure 2 illustrates how the strategies vary with S_i. As S_i increases from its minimum value of 0 to its maximum of 1, the strategies change according to the following sequence:

1. A hybrid equilibrium where the high quality individuals all signal and the low quality individuals signal with some probability that increases with their social prominence/capital: $P(r|q = 1, S) = 1$ for all S; and $0 < P^*(r|q = 0, S) < 1$ with $\frac{dP^*(r|q=0,S)}{dS} > 0$.

2. A pooling equilibrium in which all signal: $P^*(r|q = 0, S) = P^*(r|q = 1, S) = 1$ for all S.

Figure 2 also shows the change in reputation that each type of individual gains after signalling. We define this change in reputation as $\Delta\hat{q}_i = \hat{q}_i - \hat{q}_{i,t-1}$, the difference in reputation after the public signalling event compared to before the event (reputation before the event is assumed to be 0.5 for all individuals).

In the hybrid equilibrium, S_i does not affect change in reputation. Although S_i directly increases reputational gain, this effect is offset by the fact that $P^*(r|q = 0, S_i)$ increases with S_i, which reduces the reputational gain. In the pooling equilibrium in which everyone signals, signalling itself is not informative. Only the success/failure of the signal and S_i are informative. As S_i increases, it becomes more and more informative relative to the outcome, so that at high S_i, both successes and failures are thought to emanate from high quality individuals with high probability.

We thus find that the gains from engaging in costly acts that are supposed to demonstrate quality can be higher for those with higher social prominence/capital. As social prominence/capital increases, it gains in importance in the public event because it becomes the main carrier of information (since it leads to a pooling equilibrium in which the act of signalling itself is no longer informative).

Mechanism 2: altered payoff

Our second intervention is to make the payoff to the signal contingent not only on quality, but also on social prominence/capital. Individuals with higher social prominence/capital may incur lower costs from signalling, for example if they are buffered by support from others, or those with higher social prominence may accrue greater benefits, for example if information about their signalling success is broadcast more widely. To reflect this, we now consider a general payoff function $\Pi(q_i, \hat{q}_i, S_i, a_i)$ with the property that social prominence/capital increases the payoff from being perceived as high quality: $\frac{\partial^2 \Pi}{\partial S_i \partial \hat{q}_i} > 0$.

We find that at low values of S_i we have a pooling equilibrium in which no one signals. This is because the benefit of signalling is too low relative to the cost if S_i is low, even for high quality individuals (full details are presented in SI Section 1.3, with strategies plotted in Figure SI.2b). As S_i increases, we move to a fully separating equilibrium ($P^*(r|q = 0, S) = 0$, $P^*(r|q = 1, S) = 1$). Then, as S_i increases further still, we move to a hybrid equilibrium in which high quality individuals always signal and low quality individuals signal with a probability that increases with S_i, until we reach the pooling equilibrium in which all signal and observers can only distinguish $q = 1$ and $q = 0$ individuals because of the different frequency of success and failure. Thus, under some threshold S_i, observers cannot distinguish high and low quality individuals (because no one signals). Above that threshold, the capacity of observers to distinguish low and high quality individuals decreases with S_i.

In SI Figure SI.2c, we also present the strategies that emerge with the combination of these two mechanisms. It is similar to what we described for the altered payoff mechanism alone. The difference is that low quality individuals are more eager to signal, and so the probability of them signalling increases faster with S_i than when the altered payoff mechanism operates alone. This causes the pooling equilibrium in which both types signal to prevail over a larger range of S_i values. This is because the altered prior mechanism increases the payoffs from signalling for high social prominence/capital, low quality individuals. In the agent-based model below, this will affect the reputational dynamics for these individuals relative to a scenario in which only the altered payoff mechanism is at play.

3 Agent-based model: the co-evolution of reputation and social prominence/capital

We now develop an agent-based model to explore the feedbacks between social prominence/capital and the signalling behaviours analysed above. We lay out the key elements here, with the details of the model and of the simulations in SI Section 2.

In this model, we combine multiple mechanisms by which individuals attempt to learn each other's quality. First, individuals engage in pairwise interactions. Individuals "visit" each other and in so doing make an inference about the probability that the person they visit is of high quality (e.g., they notice that this person is more or less cooperative or helpful). Their observation of other's quality is noisy (see SI Section 2 for the exact way the inference is made).

Second, there are public events. We consider a baseline public event in which individuals observe each other's current level of social prominence/capital (i.e., S_i is revealed). Using S_i they infer quality $\hat{q}_i = f(S_i)$. This is the "cue only" mechanism in which individuals do not have the opportunity to perform a public signal.[3] This baseline represents the sociological tradition (e.g., Gould, 2002; Lynn et al., 2009; Manzo and Baldassarri, 2015). We then move to public events in which individuals have the opportunity to engage in a costly signal, following the strategies derived in the analytical model. As above, their social prominence/capital impacts either how the acts are interpreted (altered prior mechanism) or the payoffs they receive (altered payoff mechanism), or both simultaneously. We model different scenarios that combine these mechanisms of learning (pairwise interactions, cue only mechanism, altered prior mechanism, and altered payoff mechanism) in different ways.

Individuals are linked in a network defined by interaction weights (cf. Skyrms and Pemantle, 2000; Gould, 2002; Lynn et al., 2009; Manzo and Baldassarri, 2015). These interaction weights determine who visits whom during the pairwise interactions, and also determine S_i. These weights are updated as the individuals learn about each other's quality through the mechanisms above. Formally, the weights evolve as $w_{ij,t} = \delta w_{ij,t-1} + \hat{q}_i$ where $w_{ij,t}$ is the weight j gives to i at time t, and δ is a discount parameter (for memory decay, or present bias). That is, the weight w_{ij} accumulates information about i's observations of j's quality through the pairwise interaction and the public events. In turn, the interaction weights influence who engages in pairwise interactions, as well as each individual's social prominence/capital (which we model as a function of the sum of these weights). Hence, while S_i was exogenous in the analytical model, it is now fully endogenised.

We wish to understand the role of quality versus current social prominence/capital in shaping reputation and building up social prominence/capital over time. To do so, we introduce an initial bias in the interaction weights at the start of the simulation, which is unrelated to quality (cf. Frey and van de Rijt, 2016; Hackel and Zaki, 2018). Hence, some individuals start with a higher initial social prominence level S_i. We then consider four types of individuals: (1) individuals of high quality who start with higher S_i, (2) individuals of high quality who start with lower S_i, (3) individuals of low quality who start with higher S_i, and (4) individuals of low quality who start with lower S_i.

We present the results of four scenarios, all of which include learning through pairwise interaction, but include different learning mechanisms in the public event. These four scenarios are (1) cue only mechanism (a baseline), (2) altered prior mechanism, (3) altered payoff mechanism, and (4) altered prior and payoff combined. In SI Section **??**, we consider more scenarios, including a baseline with only private learning through pairwise interactions, and a baseline with a pure signalling game that is not influenced by social prominence/capital. Figure 3 shows individual trajectories, while Figure 4 shows the distribution of S_i across all rounds, for each type of individual.

In the Cue Only baselin model, the agent-based model produces dynamics that align with

[3] Here, it is reasonable to call this a "cue" both in the general sociological sense, and in the behavioural ecological sense, as individuals do not have any control over the revelation of their S_i.

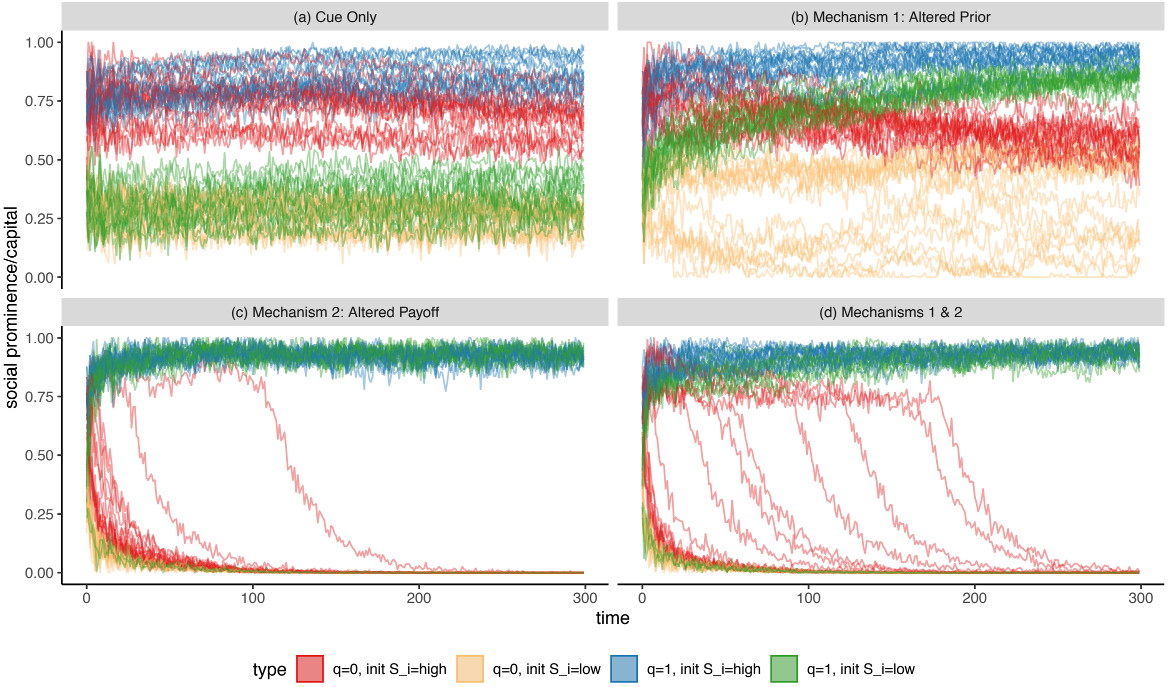


Figure 3: Representative individual time series of S_i (social prominence/capital), for individuals of different quality and initial social prominence/capital. $N = 300$, $\delta = 0.98$, $\theta_1 = 0.8$, $\theta_0 = 0.6$, $c_1 = .2$, $c_0 = .4$, and $f(S_i) = .1 + .9S_i$. The payoff function is $\Pi_i = \hat{q}_i - c_i$, except under the altered payoff mechanism when it is $\Pi_i = (S_i^2 + 0.1)\hat{q}_i - c_i$

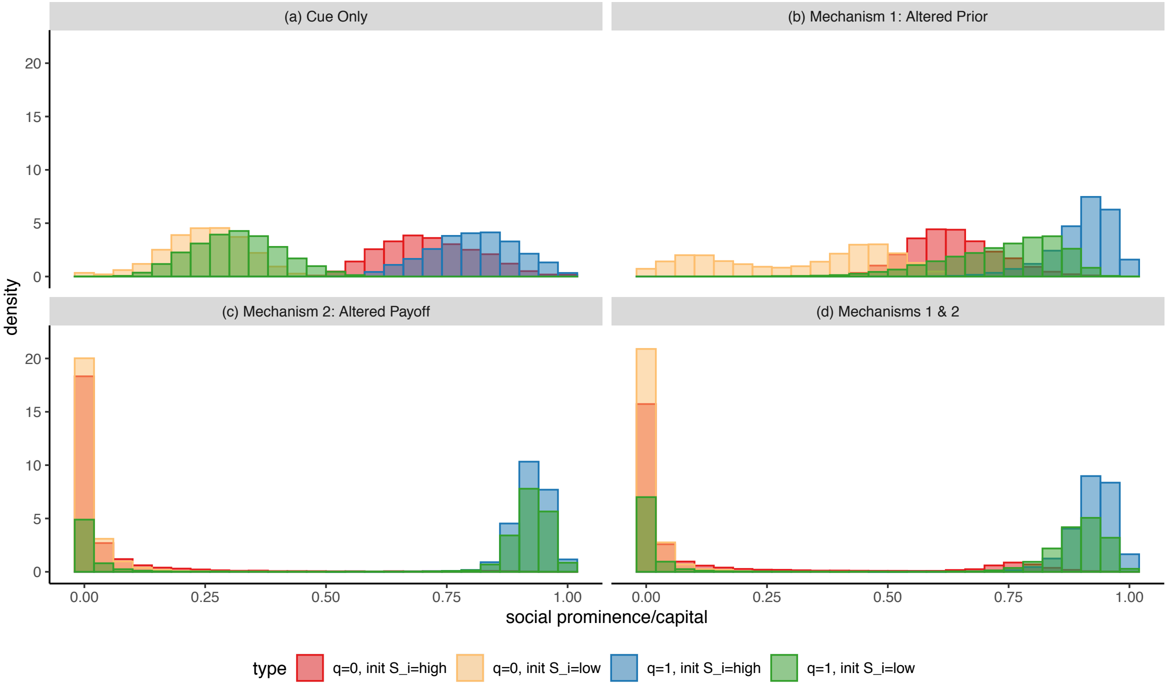


Figure 4: The distribution of S_i (social prominence/capital) in each model, as a function of the individual's quality and initial social prominence/capital. The parameters are the same as in Figure 3

the sociological work on the decoupling of status and quality researched by Gould (2002), Lynn et al. (2009), and Manzo and Baldassarri (2015). As expected, initial social prominence/capital powerfully shapes reputation. Individuals with low quality but high initial social prominence/capital are able to maintain a high level of social prominence/capital over many rounds, while individuals with high quality but low initial social prominence/capital struggle to build up their reputation (i.e., perceived quality) and social prominence/capital (see Figure 3a). Gradually, the learning that occurs via the pairwise interactions leads to declining social prominence/capital for low quality individuals and increasing social prominence/capital for high quality individuals (see Figure 3a). Yet, this process takes many rounds. As a result, quality and social prominence/capital are weakly coupled (see Figure 4a).

The SI provides a few more baselines. First, SI Section 2.2 shows that if we have only the pairwise interactions, individuals learn about each other's quality faster. Whereas if pairwise interactions are uninformative, social prominence/capital perfectly reproduces itself and stays fully decoupled from quality. Second, if individuals can signal but social prominence plays no role in that signal ("pure signalling") then types separate out quickly if the equilibrium of the signalling game is separating or hybrid (see Figure SI.4). Pure signalling fully neutralises the effect of an initial social prominence/capital advantage even though social prominence/capital shapes the pairwise interactions.

Having established these baselines, we consider the role of signalling under the altered prior mechanism. The model runs show that thanks to costly signals, observers gain useful information about quality (see Figure 3b). In particular, high quality individuals are revealed over time as such, and enjoy increasing level of social prominence/capital, independent of their starting point. In contrast, low quality individuals tend to lose reputation and social prominence/capital over time, due to the fact that they sometimes do not signal, and experience failures more frequently. This feeds back on pairwise interactions, which become less frequent (see Figure SI.5).

However, the model runs illustrate that the altered prior mechanism leads to incomplete information revelation and unequal reputational gains. Low quality individuals with high initial social prominence/capital can maintain a high level of both social prominence/capital and reputation for many rounds, despite costly signalling events; we term this a "reputational shield."

To understand this shield, remember from the analytical model that when individuals have high social prominence/capital, both types signal and information on quality is only gleaned by the rate of success or failure of the signal. Signalling is thus mostly an opportunity to reveal and advertise one's high social prominence/capital. As a result, even though low quality individuals fail more often than high quality individuals when they undertake the risky costly signal, they are shielded by their high social prominence/capital: onlookers use social prominence/capital as a prior, which favourably colours their interpretation of the signalling act. These low quality individuals, however, do show a general decline in their reputation over time, so this shield is temporary. It is more long-lived when pairwise interactions are uninformative (or do not happen), and when memory is perfect ($\delta = 1$, Figure SI.7), which allows misapprehensions to linger (see SI Section 2.3.1 and 2.3.2). We note also that the fragility of the shield stems from the fact that signals can fail with differential probability (the probability of success is such that $\theta_1 > \theta_0$). The number of successes and failures, as well as the pairwise interactions, create a gradual flow of information that slowly reveals quality. If we instead set $\theta_1 = \theta_0$, such that success and failure is now random, we find that the reputational shield is much more robust and long-lived (see SI Section 2.3.3).

Next, we turn to signalling with the altered payoff mechanism, where the payoff of signalling is contingent not just on quality, but also on social prominence/capital. Here again we find a slight "reputational shield," but it is short-lived, at best, with low quality individuals eventually experiencing a rapid "fall from grace" and losing the initial advantage of high social prominence/capital. This is because individuals are not shielded by their social prominence/capital if their signal fails, as they are with the altered prior mechanism, and as a result, signalling is less attractive for low types. They thus signal less often (hybrid equilibrium). Low quality individuals, then, are eventually identified as such and consequently lose social prominence/capital.

More notable is what happens for high quality individuals. The agent-based model shows that the consequence of the altered payoff mechanism is that it leads to a "reputational poverty trap," where high quality individuals who start with sufficiently low social prominence/standing do not engage in the public signalling event (see strategies in SI Figure SI.2b). This is best seen by observing the trajectories for $q = 1$, initial $S_i = low$ in Figure 3c. Despite being reliably revealed as high quality in the pairwise interactions, they remain stuck in an equilibrium of low social prominence/capital and low reputation because the costs of signalling are too high, greatly limiting their ability to build themselves up. This in turn limits the number of pairwise interactions they engage in, which also prevents them from building social prominence/capital. In contrast, high quality individuals who start with high social prominence/capital reap yet greater benefits from their signalling acts: they are involved in a larger proportion of the pairwise interactions, meaning that there is substantial inequality in the number of interactions that individuals have (see SI Section **??**, (cf. Anderson and Shirako, 2008; Frey and van de Rijt, 2016)). Of course, we also observe a similar pattern of relative disadvantage for high quality individuals with low initial S_i in the "cue only" condition, but it is more absolute in this scenario.

Finally, when both the altered prior and payoff mechanisms are operating simultaneously (following the strategies shown in SI Figure SI.2c) we see the continuation of the "reputational poverty trap" and a more pronounced "reputational shield" (since with the addition of the altered prior mechanism, we have added back the shielding effect of S_i in the interpretation of the signals). This reputational shield does not last indefinitely: eventually, as evidence from pairwise interactions and failed public signals accumulates, S_i decreases to the point where we reach the separating equilibrium. That equilibrium fully reveals quality, which is why we see this rapid "fall from grace." When memory is perfect ($\delta = 1$), as well as when pairwise interactions are uninformative, the reputational shield is more protracted (see SI Figures SI.4 and SI.7).

In SI Section 2.3 we show the robustness of our results to different assumptions.

4 Discussion

Making inferences about others on the basis of their actions is complicated. Generally, we should expect receivers of any signal to use the information they have at their disposal to infer the attributes and intentions of signallers. This could include the signallers past actions or, as we have explored here, their social prominence or social capital. In our models, the production of a costly signal reveals public "social information" (i.e., information gleaned from observing others). Widely used in group-living species (McGregor and Peake, 2000; Valone and Templeton, 2002; Danchin et al., 2004), social information can provide additional information that may be costly or time-consuming to acquire directly. On average, social information should be beneficial, as it can reduce uncertainty and increase the accuracy of individuals' assessments. However, our models highlight a point largely overlooked in the behavioural ecology literature (cf. Giraldeau et al., 2002), although noted by the sociological literature on status: the potential for some assessments to be less accurate, not more.

Most notably, our models show how drawing on this additional information may generally be informative, but can also result in systematic bias. In particular, we show that if social prominence colours observers' interpretation of a costly signal (by altering the observer's prior about the individual's quality), then low quality individuals with high social prominence/capital enjoy a "reputational shield" in which they "pass" as high quality for a prolonged period. Second, we show that if social prominence/capital alters the payoffs from signalling, then we can obtain a "reputational poverty trap," where high quality individuals are unable to reap the reputational benefits of their acts if they start with low social prominence/capital. Echoing other work demonstrating that signal costs may not be sufficient guarantors of signal honesty (Lachmann et al., 2001; Számadó, 2011; Fraser, 2012; Higham, 2014), we find, then, that reputation can lie.

Here, we have aimed to bridge the signalling models developed by economists and evolutionary scientists on the one hand and the status models developed by sociologists on the other. Sociologists' models of status formation importantly demonstrate the possibility of the decoupling of quality and status, as we too see in our "cue only" model. Yet, they do not give individuals the agency to undertake costly acts to reveal their underlying quality. By situating our work within the signalling theory framework, we allow for more agency on the part of individuals, with a stronger pull towards truthful revelation and interpretation because of the strategic incentives to maximise payoffs. Hence, a mechanism that shows such reputational misapprehensions within the signalling theory framework is likely to be robust to selection, learning, and strategic reasoning.

One way to understand the public act we model is as the simultaneous production of a costly signal (of quality) and an "index" (of social prominence/capital) (cf. Vehrencamp, 2000). The latter provides intrinsically reliable information, as when the roar of a red deer unfakeably indicates its body size (Maynard Smith and Harper, 1995; 2003). The public act in our models is thus a "multicomponent" or "multimodal" signal (Johnstone, 1995; Rowe, 1999; Partan and Marler, 1999; Higham and Hebets, 2013), which should generally improve reliability and transmissibility. But here again we show that in some cases it can lead to misapprehensions. Johnstone (1995) notes this possibility, but dismisses it, as the aggregate assessment will still be improved. While this may be the case, we argue that more attention should be paid to the cases where inaccurate assessments occur, and to their effects: the structural inequalities they can foster may not be inconsequential.

The second mechanism we explore, where social prominence/capital directly affects the payoffs of the signal, demonstrates these potential consequences. Our finding that disparities in initial endowments (here, of social prominence/capital) relegate some individuals to sustained and largely inescapable reputational deficit is the essence of a poverty trap, shown by economists to afflict both individuals and whole economies (Bowles et al., 2006; Ghatak, 2015). Note that this simple alteration of the payoff function not only results in this trap, but more generally increases inequality: we see a much more skewed distribution of social prominence/capital, reputation, and pairwise interactions. While signalling here may generally be reliable, the benefits of signalling fall very unevenly.

In the South Indian case, villagers differ substantially in their social capital and social prominence, and some of that variation is driven by factors beyond their control. In this context, the most obvious factors to consider are those of gender and caste, where women (Power and Ready, 2018) and Dalits (Figure 1) often have lower social prominence or social capital. Such starting disadvantages may be sufficient to jump-start the feedbacks that we explore here, reinforcing gender-, caste-, or class-based inequality (cf. O'Connor, 2019). While the details will differ in other social contexts, when certain groups suffer an initial social disadvantage, this can be amplified by signals that reveal social prominence/capital even though individuals have the agency to pro-actively "prove their worth" through costly signals. So, for example, beyond general evidence of cumulative advantage in academic citations, we also see systematic biases in citation on the basis of gender or race and ethnicity (e.g., Dworkin et al., 2020; Bertolero et al., 2020).

5 Conclusion: future directions

We have explored how signalling theory can be extended to include an individual's social prominence/capital in the decision to signal and in receivers assessments. Our focus has been less on the stable set of strategies that may be employed, and more on their consequences (cf. Frey and van de Rijt, 2016; Hackel and Zaki, 2018; Tsvetkova, 2021). We hope that this emphasis on the structural outcomes of individuals' strategic decisions will prompt more exploration in the evolutionary sciences.

Changing how (and how many) interactions take place could add new complexity to the dynamics studied here. As our model does not limit the number of pairwise interactions individuals can have, or allow individuals to refuse or select specific partners, those of higher social prominence/capital have more interactions. This means these individuals are more thoroughly assessed, while those of lower prominence/capital have fewer chances to correct any misapprehensions (but see SI Section 2.3.1). If all individuals had a larger number of interactions, the quality of all individuals might be more readily revealed; we do not currently explore how these balance. The interactions in our model also have no value beyond the information they provide; revising them to entail an exchange or game with its own payoff would add an important new dimension. Here, we have chosen to interpret our key term S_i as a proxy for either social capital or social prominence. While this agnosticism emphasises the wide applicability of our model, it also conflates two distinct concepts, which is not without risk (cf. Power and Ready, 2018). By adding more specificity to the process and nature of interactions, we may be able to establish the distinct effects of social prominence versus social capital on reputational formation and assessment.

Our model is also currently agnostic on exactly how social prominence/capital influences the payoff of signalling. Further work should investigate the form that these costs and benefits take. While costs are often assumed to be production costs directly entailed in the enactment of the signal, this need not be the case. They may, instead, be social costs imposed by receivers on "cheats" (Lachmann et al., 2001; Barker et al., 2019), exemplified by "badges of status" in sparrows (Rohwer, 1975). One finding from the altered prior mechanism is that even failed signals may be interpreted as indicating high quality if individuals have sufficiently high social prominence/capital. This implies that such individuals would not be seen as "cheats," and so would not face these socially imposed costs; instead, the burden of such costs would fall most heavily on individuals with lower social prominence/capital, further exacerbating their disadvantage (as the example of the two firewalkers also suggests). To explore this possibility, future work should explicitly explore the consequences of modelling the costs of signalling as being receiver-dependent.

An important feature of our model is that everyone observes the public signalling events (and learns aggregate information of everyone's assessments with the revelation of S_i). However, it is plausible that social network structure could impact who is able to observe whose signals (Takács et al., 2021). Relatedly, we do not extensively explore how the relative weight of private versus public information – and the extent to which these distinct information sources agree or conflict with each other – may impact the updating process (Valone, 2007) (though see SI Sections 2.2 and 2.3.1). It would be fruitful to model how widely information is aggregated, how extensively acts are observed, and how receivers balance the varied and potentially conflicting inputs they receive.

Finally, future modelling work should investigate the co-evolution of strategies with social structure, drawing on empirically-validated models of learning in strategic interactions (Camerer et al., 2003). "Quality," too, could be seen as malleable, if, for example, it is seen as embodied or human capital itself (Bourdieu, 1986; Coleman, 1988), and so may also co-evolve with an individual's social prominence/capital.

Ethics

The fieldwork was approved by the Stanford University Human Subjects Institutional Review Board.

Data and Code Access

Code and anonymised data are available at https://github.com/eapower/when-does-reputation-lie/.

Competing Interests

The authors declare no competing interests.

Funding

Funding for E.A.P.'s fieldwork was provided by a National Science Foundation Doctoral Dissertation Improvement Grant (BCS-1121326), a Fulbright-Nehru Student Researcher Award, the Stanford Center for South Asia, and Stanford University. E.A.P. and M.D. were funded by an Omidyar Fellowship, and J.L.B was funded by the Aarhus Institute of Advanced Studies - Marie Curie Cofund Fellowship.

Acknowledgements

E.A.P. thanks her interlocutors in South India for their patience and generosity and the faculty and students from the Folklore Department at Madurai Kamaraj University for their assistance. The authors are grateful for the suggestions of Michael Lachmann and the input of the attendees at the Santa Fe Institute Behavior Seminar.

References

Anderson, C. and Shirako, A. (2008). Are individuals' reputations related to their history of behavior? *Journal of Personality and Social Psychology*, 94(2):320–333.

Appadurai, A. and Breckenridge, C. A. (1976). The South Indian temple: authority, honour, and redistribution. *Contributions to Indian Sociology*, 10(2):187–211.

Barker, J. L., Power, E. A., Heap, S., Puurtinen, M., and Sosis, R. (2019). Content, cost, and context: A framework for understanding human signaling systems. *Evolutionary Anthropology: Issues, News, and Reviews*, 28(2):86–99.

Benjamin, B. A. and Podolny, J. M. (1999). Status, quality, and social order in the California wine industry. *Administrative Science Quarterly*, 44(3):563–589.

Bertolero, M. A., Dworkin, J. D., David, S. U., Lloreda, C. L., Srivastava, P., Stiso, J., Zhou, D., Dzirasa, K., Fair, D. A., Kaczkurkin, A. N., Marlin, B. J., Shohamy, D., Uddin, L. Q., Zurn, P., and Bassett, D. S. (2020). Racial and ethnic imbalance in neuroscience reference lists and intersections with gender. *bioRxiv*, page 2020.10.12.336230.

Borgerhoff Mulder, M., Bowles, S., Hertz, T., Bell, A., Beise, J., Clark, G., Fazzio, I., Gurven, M., Hill, K., Hooper, P. L., Irons, W., Kaplan, H., Leonetti, D., Low, B., Marlowe, F., McElreath, R., Naidu, S., Nolin, D., Piraino, P., Quinlan, R., Schniter, E., Sear, R., Shenk, M., Smith, E. A., von Rueden, C., and Wiessner, P. (2009). Intergenerational wealth transmission and the dynamics of inequality in small-scale societies. *Science*, 326(5953):682–688.

Bourdieu, P. (1986). The forms of capital. In Richardson, J. G., editor, *Handbook of Theory and Research for the Sociology of Education*, pages 239–258. Greenwood Press, Westport, CT.

Bowles, S., Durlauf, S. N., and Hoff, K., editors (2006). *Poverty Traps*. Princeton University Press, Princeton, NJ.

Burt, R. S. (2008). Gossip and reputation. In Lecoutre, M. and Lievre, P., editors, *Management et Réseaux Sociaux: Ressource Pour l'action Ou Outil de Gestion?*, pages 27–42. Hermès Science publications, Paris.

Camerer, C., Ho, T., and Chong, K. (2003). Models of thinking, learning, and teaching in games. *The American Economic Review*, 93(2):192–195.

Chapais, B. (1988). Rank maintenance in female Japanese macaques: experimental evidence for social dependency. *Behaviour*, 104(1):41–58.

Chase, I. D., Tovey, C., Spangler-Martin, D., and Manfredonia, M. (2002). Individual differences versus social dynamics in the formation of animal dominance hierarchies. *Proceedings of the National Academy of Sciences*, 99(8):5744–5749.

Cialdini, R. B. and Trost, M. R. (1998). Social influence: Social norms, conformity and compliance. In Gilbert, D. T., Fiske, S. T., and Lindzey, G., editors, *The handbook of social psychology*, volume 2, pages 151–192. McGraw-Hill, 4th edition.

Coleman, J. S. (1988). Social capital in the creation of human capital. *The American Journal of Sociology*, 94:S95–S120.

Correll, S. J., Ridgeway, C. L., Zuckerman, E. W., Jank, S., Jordan-Bloch, S., and Nakagawa, S. (2017). It's the conventional thought that counts: how third-order inference produces status advantage. *American Sociological Review*, 82(2):297–327.

Danchin, É., Giraldeau, L.-A., Valone, T. J., and Wagner, R. H. (2004). Public information: from nosy neighbors to cultural evolution. *Science*, 305(5683):487–491.

DiMaggio, P. and Garip, F. (2012). Network effects and social inequality. *Annual Review of Sociology*, 38(1):93–118.

DiPrete, T. A. and Eirich, G. M. (2006). Cumulative advantage as a mechanism for inequality: a review of theoretical and empirical developments. *Annual Review of Sociology*, 32:271–297.

Dworkin, J. D., Linn, K. A., Teich, E. G., Zurn, P., Shinohara, R. T., and Bassett, D. S. (2020). The extent and drivers of gender imbalance in neuroscience reference lists. *Nature Neuroscience*, 23(8):918–926.

Engh, A. L., Esch, K., Smale, L., and Holekamp, K. E. (2000). Mechanisms of maternal rank 'inheritance' in the spotted hyaena, Crocuta crocuta. *Animal Behaviour*, 60(3):323–332.

Fraser, B. (2012). Costly signalling theories: beyond the handicap principle. *Biology & Philosophy*, 27(2):263–278.

Frey, V. and van de Rijt, A. (2016). Arbitrary inequality in reputation systems. *Scientific Reports*, 6:38304.

Ghatak, M. (2015). Theories of poverty traps and anti-poverty policies. *The World Bank Economic Review*, 29(suppl_1):S77–S105.

Giraldeau, L.-A., Valone, T. J., and Templeton, J. J. (2002). Potential disadvantages of using socially acquired information. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 357(1427):1559–1566.

Godfray, H. C. J. (1991). Signalling of need by offspring to their parents. *Nature*, 352(6333):328–330.

Gould, R. V. (2002). The origins of status hierarchies: A formal theory and empirical test. *American Journal of Sociology*, 107(5):1143–1178.

Grafen, A. (1990a). Biological signals as handicaps. *Journal of Theoretical Biology*, 144(4):517–46.

Grafen, A. (1990b). Sexual selection unhandicapped by the Fisher process. *Journal of Theoretical Biology*, 144(4):473–516.

Hackel, L. M. and Zaki, J. (2018). Propagation of economic inequality through reciprocity and reputation. *Psychological Science*, 29(4):604–613.

Henrich, J. (2009). The evolution of costly displays, cooperation and religion: credibility enhancing displays and their implications for cultural evolution. *Evolution and Human Behavior*, 30(4):244–260.

Henrich, J., Chudek, M., and Boyd, R. (2015). The Big Man Mechanism: how prestige fosters cooperation and creates prosocial leaders. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1683):20150013.

Henrich, J. and Gil-White, F. J. (2001). The evolution of prestige: Freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. *Evolution and human behavior*, 22(3):165–196.

Higham, J. P. (2014). How does honest costly signaling work? *Behavioral Ecology*, 25(1):8–11.

Higham, J. P. and Hebets, E. A. (2013). An introduction to multimodal communication. *Behavioral Ecology and Sociobiology*, 67(9):1381–1388.

Hirsch, J. E. (2007). Does the H Index have predictive power? *Proceedings of the National Academy of Sciences*, 104(49):19193–19198.

Huttegger, S., Skyrms, B., Tarres, P., and Wagner, E. (2014). Some dynamics of signaling games. *Proceedings of the National Academy of Sciences*, 111(Supplement 3):10873–10880.

Huttegger, S. M., Bruner, J. P., and Zollman, K. J. S. (2015). The handicap principle is an artifact. *Philosophy of Science*, 82(5):997–1009.

Irons, W. (1979). Cultural and biological success. In Chagnon, N. A. and Irons, W., editors, *Evolutionary biology and human social behavior: an anthropological perspective*, pages 257–272. Duxbury Press, North Scituate, MA.

Johnstone, R. A. (1995). Honest advertisement of multiple qualities using multiple signals. *Journal of Theoretical Biology*, 177(1):87–94.

Lachmann, M. and Bergstrom, C. T. (1998). Signalling among relatives. II. beyond the Tower of Babel. *Theoretical Population Biology*, 54:146–160.

Lachmann, M., Számadó, S., and Bergstrom, C. T. (2001). Cost and conflict in animal signals and human language. *Proceedings of the National Academy of Sciences of the United States of America*, 98(23):13189–13194.

Lin, N. (2001). Building a network theory of social capital. In Lin, N., Cook, K. S., and Burt, R. S., editors, *Social capital: theory and research*, pages 3–30. Aldine de Gruyter, New York.

Lynn, F. B., Podolny, J. M., and Tao, L. (2009). A sociological (de) construction of the relationship between status and quality. *American Journal of Sociology*, 115(3):755–804.

Majolo, B., Lehmann, J., de Bortoli Vizioli, A., and Schino, G. (2012). Fitness-related benefits of dominance in primates. *American Journal of Physical Anthropology*, 147(4):652–660.

Manzo, G. and Baldassarri, D. (2015). Heuristics, interactions, and status hierarchies: An agent-based model of deference exchange. *Sociological Methods & Research*, 44(2):329–387.

Maynard Smith, J. and Harper, D. (1995). Animal signals: Models and terminology. *Journal of Theoretical Biology*, 177(3):305–311.

Maynard Smith, J. and Harper, D. (2003). *Animal signals.* Oxford series in ecology and evolution. Oxford University Press, Oxford; New York.

McGregor, P. K. and Peake, T. M. (2000). Communication networks: Social environments for receiving and signalling behaviour. *Acta ethologica*, 2(2):71–81.

Menger, P.-M. (1999). Artistic labor markets and careers. *Annual Review of Sociology*, 25(1):541–574.

Merton, R. K. (1968). The Matthew Effect in science. *Science*, 159(3810):56–63.

Mines, D. P. (2005). *Fierce gods: inequality, ritual, and the politics of dignity in a South Indian village.* Indiana University Press, Bloomington, IN.

Muchnik, L., Aral, S., and Taylor, S. J. (2013). Social influence bias: a randomized experiment. *Science*, 341(6146):647–651.

Newman, M. E. (2009). The first-mover advantage in scientific publication. *Europhysics Letters*, 86(6):68001.

Oakley, K. and O'Brien, D. (2016). Learning to labour unequally: Understanding the relationship between cultural production, cultural consumption and inequality. *Social Identities*, 22(5):471–486.

O'Connor, C. (2019). *The Origins of Unfairness: Social Categories and Cultural Evolution.* Oxford University Press.

Ohtsuki, H. and Iwasa, Y. (2006). The leading eight: Social norms that can maintain cooperation by indirect reciprocity. *Journal of Theoretical Biology*, 239(4):435–444.

Partan, S. R. and Marler, P. (1999). Communication goes multimodal. *Science*, 283(5406):1272–1273.

Petersen, A. M., Jung, W.-S., Yang, J.-S., and Stanley, H. E. (2011). Quantitative and empirical demonstration of the Matthew effect in a study of career longevity. *Proceedings of the National Academy of Sciences*, 108(1):18–23.

Podolny, J. M. (1993). A status-based model of market competition. *American journal of sociology*, 98(4):829–872.

Power, E. A. (2015). *Building bigness: Religious practice and social support in rural South India.* Doctoral Dissertation, Stanford University, Stanford, CA.

Power, E. A. (2017a). Discerning devotion: Testing the signaling theory of religion. *Evolution and Human Behavior*, 38(1):82–91.

Power, E. A. (2017b). Social support networks and religiosity in rural South India. *Nature Human Behaviour*, 1(3):0057.

Power, E. A. and Ready, E. (2018). Building bigness: Reputation, prominence, and social capital in rural South India. *American Anthropologist*, 120(3):444–459.

Przepiorka, W. and Diekmann, A. (2021). Parochial cooperation and the emergence of signalling norms. *Philosophical Transactions of the Royal Society B: Biological Sciences*. under review.

Ridgeway, C. L. and Erickson, K. G. (2000). Creating and spreading status beliefs. *American Journal of Sociology*, 106(3):579–615.

Roberts, G., Raihani, N., Bshary, R., Marín, M. H., Farina, A., Samu, F., and Barclay, P. (2021). The benefits of being seen to help others: indirect reciprocity and reputation-based partner choice. *Philosophical Transactions of the Royal Society B: Biological Sciences*. under review.

Rohwer, S. (1975). The social significance of avian winter plumage variability. *Evolution*, 29(4):593–610.

Rowe, C. (1999). Receiver psychology and the evolution of multicomponent signals. *Animal Behaviour*, 58(5):921–931.

Salganik, M. J., Dodds, P. S., and Watts, D. J. (2006). Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 311(5762):854–856.

Santos, F. P., Pacheco, J. M., and Santos, F. C. (2021). The complexity of human cooperation under indirect reciprocity. *Philosophical Transactions of the Royal Society B: Biological Sciences*. accepted.

Simcoe, T. S. and Waguespack, D. M. (2010). Status, quality, and attention: what's in a (missing) name? *Management Science*, 57(2):274–290.

Skyrms, B. and Pemantle, R. (2000). A dynamic model of social network formation. *Proceedings of the National Academy of Sciences of the United States of America*, 97(16):9340–9346.

Smith, E. A. (2004). Why do good hunters have higher reproductive success? *Human Nature*, 15(4):343–364.

Sommerfeld, R. D., Krambeck, H.-J., and Milinski, M. (2008). Multiple gossip statements and their effect on reputation and trustworthiness. *Proceedings of the Royal Society B: Biological Sciences*, 275(1650):2529–2536.

Sosis, R. and Alcorta, C. S. (2003). Signaling, solidarity, and the sacred: The evolution of religious behavior. *Evolutionary Anthropology*, 12(6):264–274.

Spence, M. (1973). Job market signaling. *The Quarterly Journal of Economics*, 87(3):355–374.

Számadó, S. (2011). The cost of honesty and the fallacy of the handicap principle. *Animal Behaviour*, 81(1):3–10.

Takács, K., Gross, J., Testori, M., Letina, S., Kenny, A. R., Power, E. A., and Wittek, R. P. M. (2021). Networks of reliable reputations and cooperation: a review. *Philosophical Transactions of the Royal Society B: Biological Sciences*. accepted.

Tsvetkova, M. (2021). The effects of reputation on inequality in network cooperation games. *Philosophical Transactions of the Royal Society B: Biological Sciences*. under review.

Valone, T. J. (2007). From eavesdropping on performance to copying the behavior of others: a review of public information use. *Behavioral Ecology and Sociobiology*, 62(1):1–14.

Valone, T. J. and Templeton, J. J. (2002). Public information for the assessment of quality: a widespread social phenomenon. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 357(1427):1549–1557.

van de Rijt, A., Kang, S. M., Restivo, M., and Patil, A. (2014). Field experiments of success-breeds-success dynamics. *Proceedings of the National Academy of Sciences*, 111(19):6934–6939.

Vehrencamp, S. L. (2000). Handicap, index, and conventional signal elements of bird song. In Espmark, Y., Amundsen, T., and Rosenqvist, G., editors, *Animal signals: signalling and signal design in animal communication*, pages 277–300. Tapir Academic Press.

von Rueden, C., Gurven, M., and Kaplan, H. (2011). Why do men seek status? Fitness payoffs to dominance and prestige. *Proceedings of the Royal Society B: Biological Sciences*, 278(1715):2223 –2232.

von Rueden, C. R., Trumble, B. C., Thompson, M. E., Stieglitz, J., Hooper, P. L., Blackwell, A. D., Kaplan, H. S., and Gurven, M. (2014). Political influence associates with cortisol and health among egalitarian forager-farmers. *Evolution, Medicine, and Public Health*, 2014(1):122–133.

Wagner, E. O. (2013). The dynamics of costly signaling. *Games*, 4(2):163–181.

Whitmeyer, M. (2021). Strategic inattention in the Sir Philip Sidney Game. *Journal of Theoretical Biology*, 509:110513.

Xygalatas, D., Maňo, P., Bahna, V., Kundtová-Klocová, E., Kundt, R., Lang, M., and Shaver, J. H. (2021). Social inequality and signaling in a costly ritual. *Evolution and Human Behavior*. accepted.

Zollman, K. J. S., Bergstrom, C. T., and Huttegger, S. M. (2012). Between cheap and costly signals: the evolution of partially honest communication. *Proceedings of the Royal Society B: Biological Sciences*.

Supporting Information: When does reputation lie? Dynamic feedbacks between costly signals, social capital, and social prominence

Marion Dumas
Grantham Research Institute, London School of Economics & Political Science

Jessica L. Barker
Aarhus University Interacting Minds Centre, Alaska Dept of Health & Social Services

Eleanor A. Power
London School of Economics & Political Science

July 19, 2021

Contents

1 Analytical model derivations

See Figure SI.1 for a schematic representation of the key elements for the altered prior and altered payoff mechanisms.

1.1 Equilibria of the signalling games

We reiterate the elements of the game:

- Ingredients:
  - N individuals each with quality $q_i \in \{0, 1\}$
  - Common priors about q_i, denoted π_{1i}. Note that in the altered prior mechanism, S_i affects this prior.
  - Social prominence/capital S_i, set to 0 for the individual with lowest prominence/capital and 1 for the highest.
  - Action space $a \in \{r, \neg r\}$ (to signal or not to signal).
  - The outcome $o \in \{s, \neg s\}$ of each signal is probabilistic: a signal succeeds ($o = s$) with probability θ_1 for individuals with quality $q_i = 1$ and with probability θ_0 for individuals with quality $q_i = 0$.
  - Payoffs are $P_i(q_i, \hat{q}_i, a_i)$, where $\hat{q}_i$ is the inference made about i's quality at the end of the game. The payoff function has the following properties: 1) $\Pi(1, \hat{q}_i, r) - \Pi(1, \hat{q}_i, \neg r) > \Pi(0, \hat{q}_i, r) - \Pi(0, \hat{q}_i, \neg r)$, i.e. it is costlier to signal for type $q = 0$, and 2) it is beneficial to be perceived to be of high quality, which is expressed as $\frac{d\Pi(q_i, \hat{q}_i, a_i)}{d\hat{q}_i} > 0$. Note that the in altered payoff mechanism, S_i also enters the payoff function, but the two properties above continue to hold.
- The game unfolds as follows:
  - All players simultaneously decide on their signalling strategy $P(r|q_i, S_i)$.
  - For players with mixed strategies $0 < P(r|q_i, S_i) < 1$, a move of nature decides a_i, i.e. whether they signal, according to probability $P(r|q_i, S_i)$.
  - For each player i who signals, a move of nature decides o_i (whether her signal is successful).
  - All players observe a_i and o_i for all other players, and make the same inference $\hat{q}_i | a_i, o_i$ using Bayes' formula.
  - Payoffs are realized.

The solution concept we use to solve this static costly signalling game is that of a Bayesian Nash equilibrium. A Nash equilibrium is a strategy profile (the combination of each type's strategy for a given value of S: $(P(r|q = 1, S), P(r|q = 0, S))$ in which neither player would change strategy given the strategy of the other player. A Bayesian Nash equilibrium puts a constraint on the beliefs that actors hold. Because we are in a game of incomplete information (the types of other actors are unknown), actors hold probabilistic beliefs about the type of other actors given the actions they observe. In a Bayesian Nash equilibrium, these beliefs are consistent with the strategies of actors and they are updated by Bayes' rule after observing actors' actions. An example of a violation in our system would be if an observer interpreted

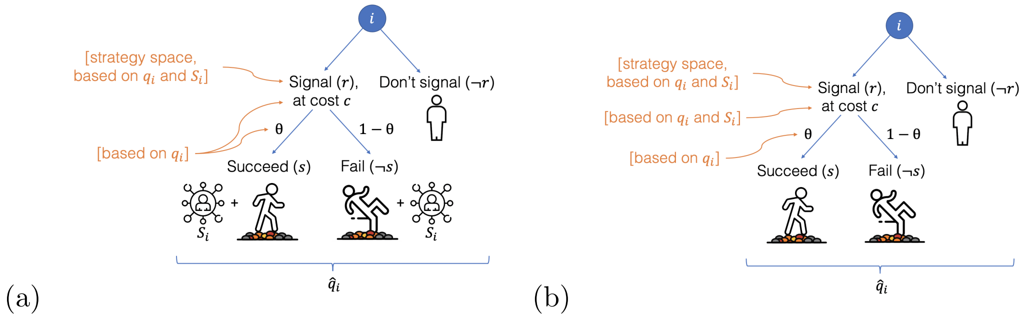


Figure SI.1: Schematic of the analytical models. Individual i faces the choice to signal (r) or not ($\neg r$). Signalling entails a cost (c_0 for $q = 0$ and c_1 for $q = 1$, and succeeds with probability θ_0 for $q = 0$ individuals and θ_1 for $q = 1$ individuals). S_i (social prominence/capital) enters in either by being revealed during the public signalling event and influencing observers' inferences about i's quality $\hat{q}_i$ (a, altered prior mechanism), or by effectively altering the cost c of signalling (b, altered payoff mechanism). [Image credits: pongsakornRed, Freepik, Smashicons, Kiranshastry].

a signaller as being type 0 (low quality) even though the strategy of a type 0 actor is to never signal.

Denote as (P_0^*, P_1^*) the strategy profiles $(P(r|q = 1, S), P(r|q = 0, S))$ which can be equilibria. These can be:

- $(0, 0)$: pooling equilibrium with no signal
- $(0, 1)$: separating equilibrium where type 1 individuals signal and type 0 individuals do not.
- $(0 < P_0^* < 1, 1)$: hybrid equilibrium where type 0 signal with some probability and type 1 signal always. This is the case when type 0 is indifferent between signalling and not signalling.
- $(1, 1)$: pooling equilibrium where all signal.

Note that $(0 < P_0^* < 1, 0)$ and $(1, 0 < P_1^* < 1)$ are trivially excluded because the cost of signalling is higher for low quality individuals. $(0 < P_0^* < 1, 0 < P_1^* < 1)$ is excluded because it is not possible for both type 1 and type 0 actors to be simultaneously indifferent between r and $\neg r$. Finally, $(0, 0 < P_1^* < 1)$ is also excluded because when such an equilibrium exists, $(0, 1)$ exists as well, so we focus on $(0, 1)$ to simplify our analysis (none of our important conclusions are affected by this).

The strategy profile must be solved for every value of $S_i \in [0, 1]$ since S_i affects the payoffs either directly (altered payoff mechanism) or via $\hat{q}_i$ (altered prior mechanism). Hence, which of the above strategy profiles can be an equilibrium will depend on S_i. In what follows, we outline the conditions that must be satisfied for each of the four types of strategy profiles to constitute an equilibrium. These conditions hold independently of whether we are considering the altered prior or altered payoff mechanism, so we include S_i as an argument of the payoff function to capture both cases. In all the equations below, $\hat{q}$ is the inferred quality given the strategies P_0^* and P_1^*, and given the outcome s and $\neg s$. The inferred quality follows Equations

1, 2 and 3 (main text), except when the action is off-equilibrium. In these cases, we specify below the belief formed given an off-equilibrium action.

The condition for a pooling equilibrium with no signal $(0, 0)$ is:

$$\Pi(1, \hat{q}_i, S_i, r) < \Pi(1, \hat{q}_i, S_i, \neg r) \Rightarrow \Pi(1, 1, S_i, r) < \Pi(1, 0.5, S_i, \neg r) \tag{SI.1}$$

In this equilibrium, signalling is off-equilibrium, so we assume that if a signal is made, observers infer that the signaller is type 1. Whereas if no one signals, $\hat{q} = 0.5$ for everyone. Note too that whenever signalling is fully informative about type (such as in this off-equilibrium signal), the outcome (s or $\neg s$) doesn't matter and can be ignored.

Equation SI.1 says that for $(0, 0)$ to be an equilibrium, type 1 must prefer not to signal and have reputation $\pi_{1i} = 0.5$ in the eyes of the observers, than to signal and obtain a reputation $\hat{q} = 1$ in the eyes of observers. When this inequality is satisfied, then it is automatically true that type 0 individuals also do not signal (since their cost is higher, the equivalent inequality for type 0 is automatically satisfied).

The conditions for a separating equilibrium $(0, 1)$ are:

$$\Pi(1, \hat{q}, S_i, \neg r) < \Pi(1, \hat{q}, S_i, r) \Rightarrow \Pi(1, 0, S_i, \neg r) < \Pi(1, 1, S_i, r) \tag{SI.2}$$

$$\Pi(0, \hat{q}, S_i, r) < \Pi(0, \hat{q}, S_i, \neg r) \Rightarrow \Pi(0, 1, S_i, r) < \Pi(0, 0, S_i, \neg r) \tag{SI.3}$$

In this equilibrium, those who signal are inferred to be type 1 ($\hat{q}|r = 1$). The outcome (s or $\neg s$) is irrelevant since the signal itself is fully informative in a separating equilibrium. Those who do not signal are inferred to be type 0 ($\hat{q}|\neg r = 0$). Hence, $(0, 1)$ is an equilibrium if it is worth it for type 1 to signal and obtain reputation 1 and not worth it for type 0 to signal, even though this means he/she will get reputation 0.

The conditions for a hybrid equilibrium are:

$$\Pi(1, 0, S_i, \neg r) < E[\Pi(1, \hat{q}, S_i, r)] \tag{SI.4}$$

$$\Pi(0, \hat{q}, S_i, \neg r) = E[\Pi(0, \hat{q}, S_i, r)] \tag{SI.5}$$

In this equilibrium, the decision to participate in the public signal is not fully informative (since $0 < P_0^*(S_i) < 1$, so the observer cannot be sure that an actor signalling is type 1 and an actor not signalling is type 0). Hence the outcome (s or $\neg s$) of the signal is informative (as long as $\theta_0 \neq \theta_1$) and will affect $\hat{q}$. Since the outcome is stochastic, the actors must consider their expected payoff from signalling, where the expectation is taken over the outcomes $o \in \{s, \neg s\}$.

In this equilibrium, type 0 is indifferent between signalling and not signalling. Specifically there exists $P_0^*(S_i)$ such that condition SI.5 holds. When this condition holds, then type 1 individuals automatically prefer to signal, as expressed by inequality SI.4.

Finally, the condition for the pooling equilibrium where all signal $(1, 1)$ is:

$$\Pi(0, 0, S_i, \neg r) < E[\Pi(0, \hat{q}, S_i, r)] \tag{SI.6}$$

Type 0 must prefer to signal than not to signal and get reputation 0 (in this case, not signalling is an off-equilibrium action; we assume that observers who do not signal when everyone else does are inferred to be type 0). Again, in this equilibrium, outcomes are informative since signalling is not, and so we take the expectation over outcomes $o \in \{s, \neg s\}$.

With these conditions, we can study which strategy profile can be an equilibrium at different values of S_i, and how P_0^* will depend on S_i in the hybrid equilibrium. This will depend on how S_i affects the payoffs.

We now consider each mechanism in turn.

1.2 Mechanism 1: altered prior

With the altered prior mechanism, S_i serves as an indicator of quality, such that the prior is $\pi_{1i} = f(S_i)$ instead of $\pi_{1i} = .5$ ($f'(S_i) > 0$) for those who signal. Thus the equations for the inference of quality are (modified from Equations 1,2 and 3):

$$\hat{q}_i(S_i)|r, s = \frac{\theta_1 P(r|q=1) f(S_i)}{\theta_1 P(r|q=1) f(S_i) + \theta_0 P(r|q=0)(1-f(S_i))} \tag{SI.7}$$

$$\hat{q}_i(S_i)|r, \neg s = \frac{(1-\theta_1) P(r|q=1) f(S_i)}{(1-\theta_1) P(r|q=1) f(S_i) + (1-\theta_0) P(r|q=0)(1-f(S_i))} \tag{SI.8}$$

$$\hat{q}_i|\neg r = \frac{(1-P(r|q=1))}{(1-P(r|q=1)) + (1-P(r|q=0))} \tag{SI.9}$$

Here we assume that $\Pi(1,1,r) > 0$ and $\Pi(0,1,r) > 0$: a low quality individual prefers to be perceived as high quality even if she incurs the cost of signalling. Thus we cannot have a separating equilibrium. Hence, we either have a hybrid equilibrium with $P_0^* < 1$ or a pooling equilibrium with $P_0^* = P_1^* = 1$.

We first establish that in the hybrid equilibrium, P_0^* increases with S_i. Let us suppose that for some social prominence/capital level S_i we have a hybrid equilibrium. Then we know that there exists P_0^* such that:

$$E[\Pi(0, \hat{q}(S_i), r)] = \Pi(0, \hat{q}, \neg r) \tag{SI.10}$$

Equation SI.10 defines the equilibrium value P_0^* implicitly. We thus use implicit differentiation to determine how it varies with S_i.

$$\frac{DE[\Pi(0, \hat{q}(S_i), r)]}{DS_i} = \underbrace{\frac{d\Pi(0, \hat{q}, \neg r)}{dS_i}}_{=0}$$

The right hand side is equal to 0 because the payoff function does not directly include S_i and Equation SI.9 does not depend on S_i either. The above then implies:

$$\begin{aligned} \frac{DE[\Pi(0, \hat{q}(S_i), r)]}{DS_i} &= \frac{dE[\Pi(0, \hat{q}(S_i), r]}{d\hat{q}} \frac{D\hat{q}}{DS_i} = 0 \\ &= \frac{dE[\Pi(0, \hat{q}(S_i), r]}{d\hat{q}} \left(\frac{d\hat{q}}{df} \frac{df}{dS_i} + \frac{d\hat{q}}{dP_0^*} \frac{dP_0^*}{dS_i} \right) = 0 \\ &\Rightarrow \underbrace{\frac{d\hat{q}}{df} \frac{df}{dS_i}}_{>0} = - \underbrace{\frac{d\hat{q}}{dP_0^*}}_{<0} \frac{dP_0^*}{dS_i} \end{aligned} \tag{SI.11}$$

The left hand side is positive because with the altered prior mechanism, social prominence/capital acts as an indicator that informs the prior ($f(S_i)$), boosting $\hat{q}$. On the right hand side, $\hat{q}$ decreases as P_0^* increases since the signal becomes less useful for identifying high quality individuals as more low quality individuals participate. Hence, for the equality to hold, $dP_0^* dS_i$ must also be positive in the hybrid equilibrium. At some value of S_i, P_0^* increases to 1 and we get to the pooling equilibrium in which all signal.

Having established that P_0^* increases with S_i in the hybrid equilibrium, let us examine how reputational gain changes with S_i. Let us start with the hybrid equilibrium first. Equation SI.11 shows that $\frac{D\hat{q}}{DS_i} = 0$ in the hybrid equilibrium. In contrast, for all S_i values for which pooling is the equilibrium, P_0^* and P_1^* are constant. Therefore $\frac{d\hat{q}}{dS_i} = \frac{d\hat{q}}{df}\frac{df}{dS_i} > 0$. If S_i is unknown before the public event, then reputational gain is simply $\hat{q}(S_i) - 0.5$. Indeed, without any information, $\pi_{1i} = 0.5$ if we assume there are as many low quality as high quality individuals. Thus reputational gain is null in the hybrid equilibrium and increasing with S_i in the pooling equilibrium.

Thus, to explain the pattern in Figure 1, the public signalling event must reveal unknown information about social prominence or social capital. Indeed, if S_i is already fully known before the event, reputational gain is $\hat{q}(S_i) - f(S_i)$. This would clearly *decrease* as a function of S_i in the hybrid equilibrium. In the pooling equilibrium it would also decrease: $\frac{\Delta\hat{q}}{dS_i} = \frac{df}{dS_i}(\frac{d\hat{q}}{df} - 1) < 0$

1.3 Mechanism 2: altered payoff

Figure SI.2b qualitatively depicts the succession of equilibria under the altered payoff mechanism as the social prominence/capital of individuals increases. The following explains why equilibria succeed each other in this way.

S_i now affects the payoff directly by increasing the extent to which the signaller's reputation is broadcast. This means that S_i and $\hat{q}$ are complementary: $\frac{d^2\Pi(0,\hat{q},S_i,r)}{dS_i d\hat{q}} > 0$ (it is for example satisfied if $\Pi = g(S_i)\hat{q} - c_i$).

Consider individuals with a specific social prominence/capital level S_n such that $(0, 0)$ is an equilibrium for these individuals and so inequality SI.1 holds. This happens if S_n is very low and thus lowers the value of having a high reputation $\hat{q}_i$ even for the high individual whose cost of signalling is c_1.

Now consider another individual with social prominence/capital level $S_p > S_n$. Since prominence and reputation are complementary, at some value S_p large enough, the inequality SI.1 no longer holds. At that point, $\Pi(1, 0.5, \neg r, S_i) < \Pi(1, 1, r, S_i)$. This satisfies the inequality SI.2, one of the conditions for the separating equilibrium. This separating equilibrium will be stable for all $S_i \geq S_p$ as long as $\Pi(0, 1, r, S_i) < \Pi(0, 0, \neg r, S_i)$ (inequality SI.3).

Now consider $S_q > S_p$. Because of the complementarity between S_i and $\hat{q}$, inequality SI.2 remains true as S_i increases. However, $\Pi(0, 1, r, S_i)$ increases with S_i, while $\Pi(0, 0, \neg r, S_i)$ decreases (or stays constant) with S_i. Hence, there is some $S_q > S_p$ at which SI.3 no longer holds and type 0 individuals will choose to signal with positive probability ($P_0^* > 0$).

We have just established that as S_i increases, we move from a pooling equilibrium, to a separating equilibrium, to a hybrid equilibrium ($P(r|q = 0) > 0, 1$). In hybrid equilibria, individuals of type 0 signal with some probability $P_0^* > 0$, determined by condition SI.5.

We now show that the value of P_0^* which satisfies condition SI.5 increases with S_i, as shown in Figure SI.2b.

Condition SI.5 defines the equilibrium value of P_0^* implicitly, and so we use implicit differentiation to determine how it varies with S_i.

$$\frac{d\Pi(0,0,S_i,\neg r)}{dS_i} = \frac{dE[\Pi(0,\hat{q},S_i,r)]}{dS_i} + \frac{dE[\Pi(0,\hat{q},S_i,r)]}{dP_0^*}\frac{dP_0^*}{dS_i}$$

$$\frac{dP_0^*}{dS_i} = \left(\frac{d\Pi(0,0,S_i,\neg r)}{dS_i} - \frac{dE[\Pi(0,\hat{q},S_i,r)]}{dS_i}\right)\Bigg/\frac{dE[\Pi(0,\hat{q},S_i,r)]}{dP_0^*}$$

$\hat{q}_i$ in Equations 1 and 2 (main text) decreases with $P(r|q=0)$ (the inference about quality from signalling is lowered if more low quality individuals participate), hence the denominator is negative. Furthermore, because of the complementarity of S_i and $\hat{q}$, the numerator is also negative. Thus, we have that $\frac{dP_0^*}{dS_i} > 0$.

We have established that P_0^* increases with S_i. At some level S_i, $P_0^* = 1$, and we have the pooling equilibrium in which all individuals signal and inequality SI.6 holds. Note, however, that when all signal, signalling in and of itself loses its information content (since all types do it), although the outcome from the signal is still informative (since θ_1 and θ_0 differ).

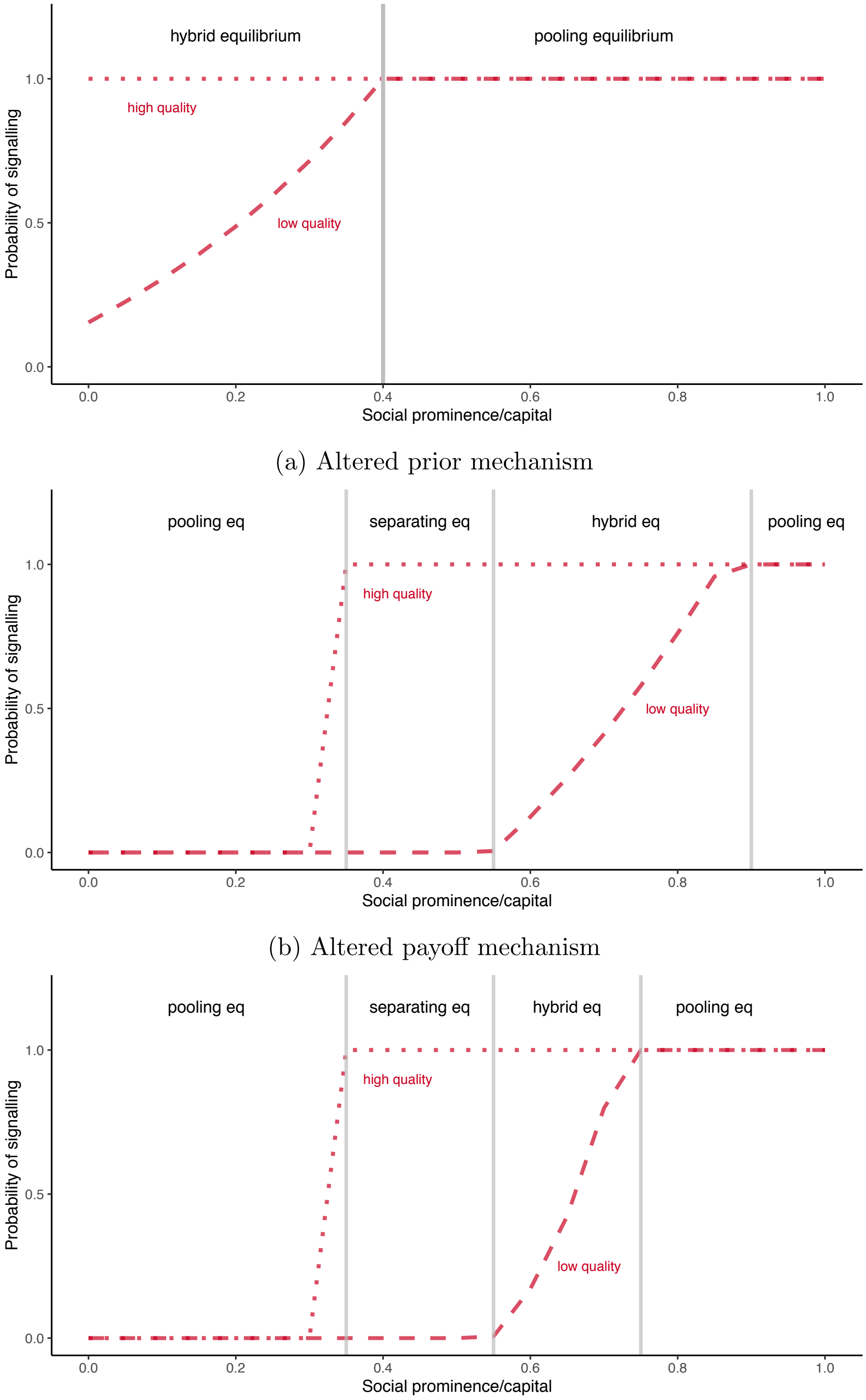
Nonetheless, as $(1,1)$ becomes a viable equilibrium at high S_i, $(0,0)$ can simultaneously be another viable equilibrium. Indeed, inequalities SI.1 and SI.6 can hold simultaneously. Additional norms outside of the scope of this model would determine which of these two equilibria prevails at a given S_i.

With the altered payoff mechanism, $\hat{q}$ is affected by S_i only through S_i's effect on P_0^*, hence $\frac{d\hat{q}}{dS_i} = \underbrace{\frac{d\hat{q}}{dP_0^*}}_{<0}\underbrace{\frac{dP_0^*}{dS_i}}_{>0} < 0$. This shows that the altered payoff mechanism cannot alone account for the pattern that reputational gain often increases with S_i (Figure 1).

1.4 Figures illustrating each mechanism and their combination

Figure SI.2a illustrates the strategies for the altered prior mechanism (equivalent to what is shown in Figure 2 in the main text), Figure SI.2b illustrates the strategies for the altered payoff mechanism alone, and Figure SI.2c illustrates the strategies for both the altered prior and altered payoff mechanisms taken together. The parameters and functional form used here are consistent with those used in Figure 2 in the main text: $\theta_1 = .8$, $\theta_0 = .6$, $c_1 = .2$, $c_0 = .4$. The payoff function in the absence of the altered payoff mechanism is simply $\Pi_i = \hat{q}_i - c_i$. The payoff function with the altered payoff mechanism is $\Pi_i = (S_i^2 + 0.1)\hat{q}_i - c_i$. The prior function $f(S_i)$ (for the altered prior mechanism) is $f(S_i) = 0.1 + .9S_i$ (so that the priors range between 0.1 and 0.9). These are also the parameters and functional forms used in Figures 3 and 4 and in fact in most figures of Section 2 unless otherwise specified.

We see the succession of the four equilibrium strategy regimes in the case of the altered payoff mechanism, with a pooling equilibrium where no one signals at low values of S_i. When combining the altered prior and altered payoff mechanisms, we see that the pooling equilibrium where all individuals signal arises earlier. This is because the influence of S_i on the prior increases the benefits of signalling even for low quality individuals and shields them from losing reputation in case of a failure.



(a) Altered prior mechanism

(b) Altered payoff mechanism

(c) Altered prior and altered payoff mechanisms combined

Figure SI.2: Equilibrium strategies of the altered prior mechanism, the altered payoff mechanism, and the altered prior and altered payoff mechanisms in combination.

2 Agent based model

See Figure SI.3 for a schematic representation of the key elements of the agent based model.

2.1 Model structure in detail

Initialisation

- Derive the strategies of the signalling interaction for each type of individual and each level of social prominence/capital $S_i = 0, 0.01, 0.02, ..., 1$. The parameters that govern these strategies are the parameters governing the signalling act itself: θ_1, θ_0, (the probabilities of signal success) c_1, c_0 (the costs of the signal), and the parameters governing how S_i affects the payoff function and the prior perception of quality π_{1i}. In what follows, $\Pi_i|r = (S_i^2 + 0.1)\hat{q}_i - c_i$ and $\pi_{1i} = .1 + .8 * S_i$.

- Initialise $N = 300$ individuals. For each:
  - draw at random its quality $q = 0$ or $q = 1$
  - draw at random whether the individual enjoys an initial favourable bias in her social prominence/capital (bias = 1 or bias = 0).

- Initialise the interaction weights $w_{ij,t=0}$ in the following way ($w_{ij,t}$ captures the relative propensity of j interacting with i (in the case of the pairwise interactions) or supporting i (in the case of show of social support during the public event) at time t as specified in detail below) :
  - If i has bias = 1, randomly draw a vector of 299 weights $w_{ij,t=0}$ from $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$
  - If i has bias = 0, randomly draw a vector of 299 weights $w_{ij,t=0}$ from $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$

  We thus obtain a weighted adjacency matrix in which some individuals have a slight initial advantage in the number of weights they get from other individuals. Introducing this initial bias allows us to study the role of quality versus social prominence/capital in the reputational and social trajectories of each agent.

Dynamic process:

1. Pairwise interactions: at each time step t,
   - For each individual j, we draw M individuals whom she will visit in this time step. The draws are done according to a multinomial distribution with probability vector given by $\frac{w_{ij,t}}{\sum_{j \neq i} w_{ij,t}}$
   - Each visit by j to each i among the M people drawn above allows j to learn about i's quality with some noise. More precisely, during her visit, j experiences the quality of the interaction with i (e.g., a more cooperative attitude on the part of i). The quality of the experience j has when interacting with i is a random variable $V_t \sim N(q_i, \sigma)$. This random variable is thus a noisy observation of i's quality. This experience allows j to draw an inference about i's quality using Bayes' formula: $\hat{q}^B_{i,t} = \frac{f(V_t; q=1, \sigma)}{f(V_t; q=0, \sigma) + f(V; q=1, \sigma)}$
   - After each visit, the weight w_{ij} is updated to reflect the inference: $w_{ij,t+1} = \delta w_{ij,t} + \hat{q}^B_{i,t}$, where δ is a discount factor to capture memory, or present bias.

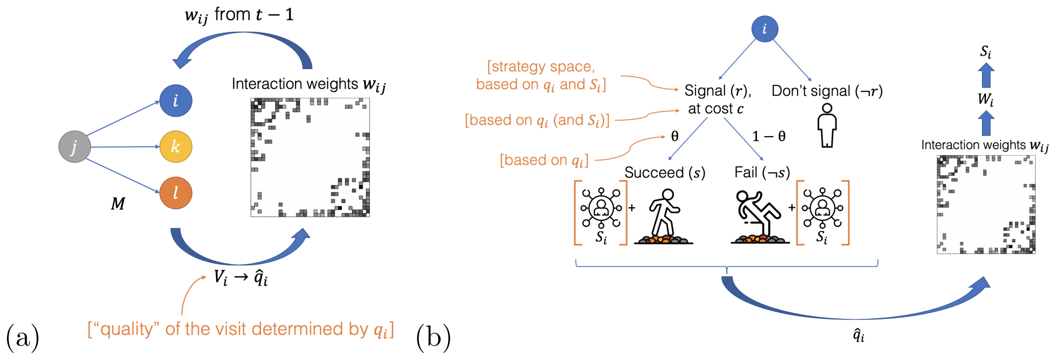

w_{ij} from $t-1$
Interaction weights w_{ij}
i
j
k
l
M
$V_i \rightarrow \hat{q}_i$
["quality" of the visit determined by q_i]
(a)
(b)
i
[strategy space, based on q_i and S_i]
Signal (r), at cost c
Don't signal ($\neg r$)
[based on q_i (and S_i)]
[based on q_i]
θ
1 − θ
Succeed (s)
Fail ($\neg s$)
S_i
+
S_i
W_i
Interaction weights w_{ij}
$\hat{q}_i$

Figure SI.3: Schematic of the agent based models, showing the pairwise interactions (a) and subsequent public signalling event (b). [Image credits: pongsakornRed, Freepik, Smashicons, Kiranshastry].

2. Public event: at the end of each time step t, after the pairwise interactions, there is a public event unfolding as follows:

   - Compute S_i, each individual's level of social prominence/capital: Each individual i may enjoy a certain level of social prominence/capital S_i (e.g., through signs of deference or accolades). This is determined by the sum of the interaction weights going to i: $W_i = \sum_{j \neq i} w_{ij}$. Social prominence/capital S_i is then the normalisation of W_i between 0 and 1 according to $S_i = \frac{W_i - min(\mathbf{W})}{max(\mathbf{W}) - min(\mathbf{W})}$ (where $\mathbf{W}$ is simply the vector of all W_i).
   - Cue only scenario: there is no opportunity for costly signalling, just an aggregation of information about social prominence/capital. In this scenario, everyone observes S_i for each individual. They use S_i as an indicator of the quality of i: $\pi_{1i} = .1 + .8 * S_i$. The weights are again updated $w_{ij,t+1} = \delta w_{ij,t} + \pi_{1i}$
   - costly signalling scenario: individuals can engage in costly signalling. Each individual determines whether to signal or not according to their S_i and quality, based on the Nash equilibrium strategies $P(r|q_i, S_i)$. If we are in a hybrid equilibrium with low quality individuals using a mixing strategy $0 < P(r|q_i, S_i) < 1$, then their decision is a random draw according to that probability. Then, for those who signal, the success or failure of their signal is drawn at random using the probabilities θ_1 and θ_0, according to their respective qualities.
   - Inferences $\hat{q}_i$ are drawn according to Equations 1, 2 and 3 (main text), and weights are again updated to $w_{ij,t+1} = \delta w_{ij,t} + \hat{q}_i$.

2.2 Outcomes under alternative combinations of mechanisms

The agent based model combines in a modular way multiple mechanisms by which individuals attempt to learn about each others' quality (private learning through pairwise interactions, social cues, signalling with the altered prior mechanism, and signalling with the altered payoff mechanism). The different possible combinations are:

1. Private learning (through pairwise interactions) only

2. Social prominence/capital as a cue only (no private learning)
3. Private learning, and a signalling event where social prominence/capital plays no role. Here we consider parameters that lead to a pooling equilibrium (where both types signal, but still with different probability of success) and parameters that lead to a hybrid equilibrium (where the low quality individual signals with some probability).
4. Signalling with the altered prior mechanism only without private learning
5. Signalling with the altered payoff mechanism only without private learning
6. Signalling with the altered prior and altered payoff mechanisms combined without private learning
7. Private learning and social prominence/capital as a cue (in the main text)
8. Private learning and signalling with the altered prior mechanism (in the main text)
9. Private learning and signalling with the altered payoff mechanism (in the main text)
10. Private learning and signalling with the altered prior and altered payoff mechanisms combined (in the main text)

In the main text, we present the last 4 scenarios. To provide baselines, we show the outcomes for the first 6 scenarios here in Figure SI.4.

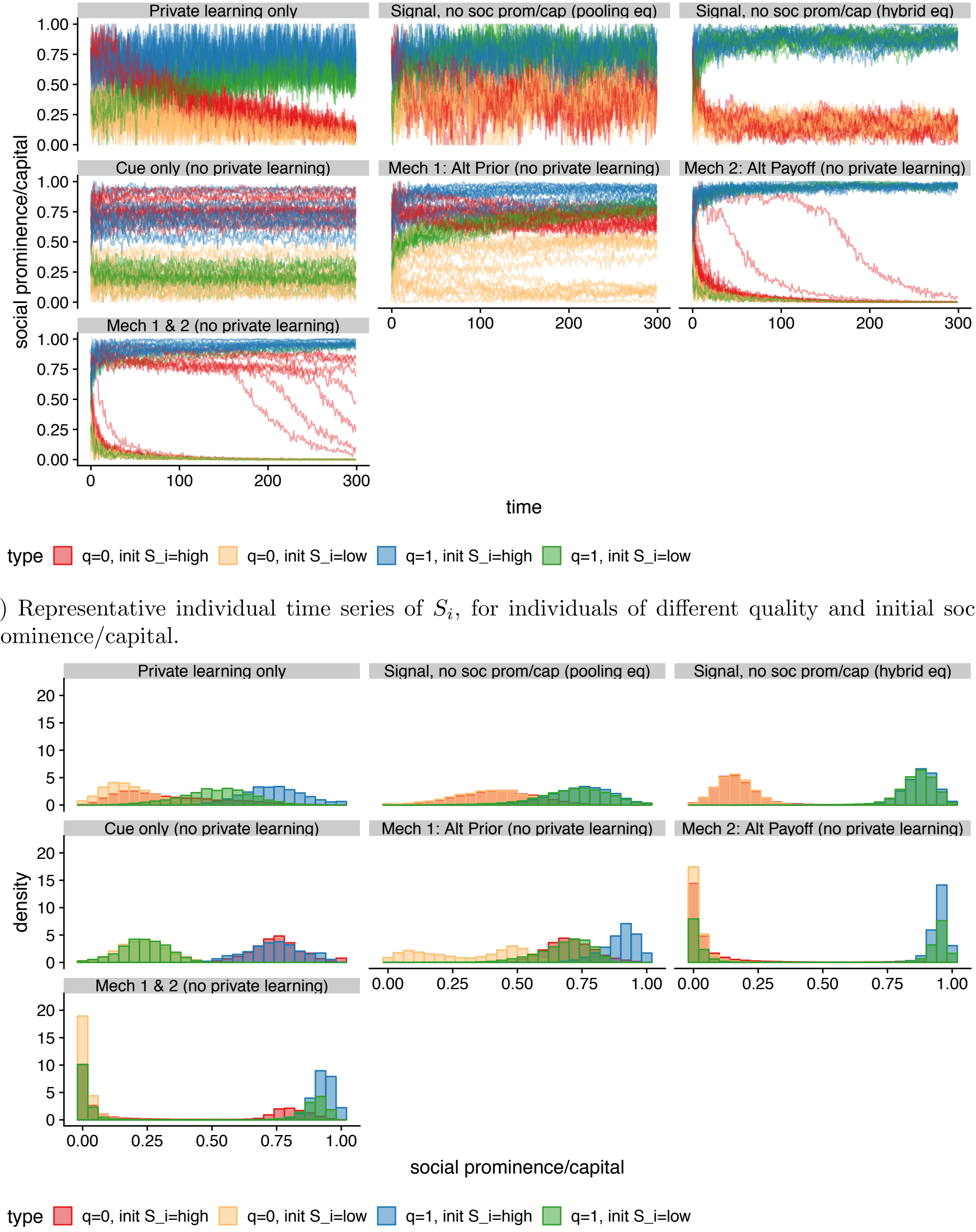
Private learning causes a slow learning about true quality, reflected in the trajectories of S_i. This leads to a distribution of social prominence/capital that reflects true quality (although high quality individuals with low initial prominence/capital maintain a long-term disadvantage as they on average receive fewer pairwise interactions to prove their worth).

Private learning with signalling (but no effect of social prominence/capital) can accelerate the learning about quality if the strategies are hybrid equilibrium or fully separating (third scenario represented in Figure **??**. If the parameters are such that the signalling equilibrium is a pooling equilibrium in which everyone signals, then learning happens through the pairwise interactions and observations about the success and failure of the signals. We see that the initial social prominence/capital of an individual plays no lingering role in this scenario and that low quality individuals are able to maintain high social prominence/capital, albeit lower than in the case of the hybrid equilibrium.

In contrast, in the model that features only the cue of social prominence/capital (without private learning), we have the opposite outcome. No one learns about quality (reflected by the fact that the trajectories do not show any temporal trend) and the distribution of social prominence/capital reflects initial differences in social prominence/capital.

Signalling with the altered prior mechanism in the absence of private learning still allows for some learning. Yet this is less complete than in Figure 3 and 4 of the main text. Indeed, low quality and high initial social prominence/capital individuals are in the "pooling equilibrium" in which everyone signals (the threshold between hybrid and pooling is around $S_i = 0.4$). Learning only occurs because of the difference in the frequency of signal success between high and low quality individuals.

Signalling with the altered payoff mechanism in the absence of private learning leads to similar patterns as with private learning: the signalling equilibrium separates types, except for those stuck in the reputational poverty trap.



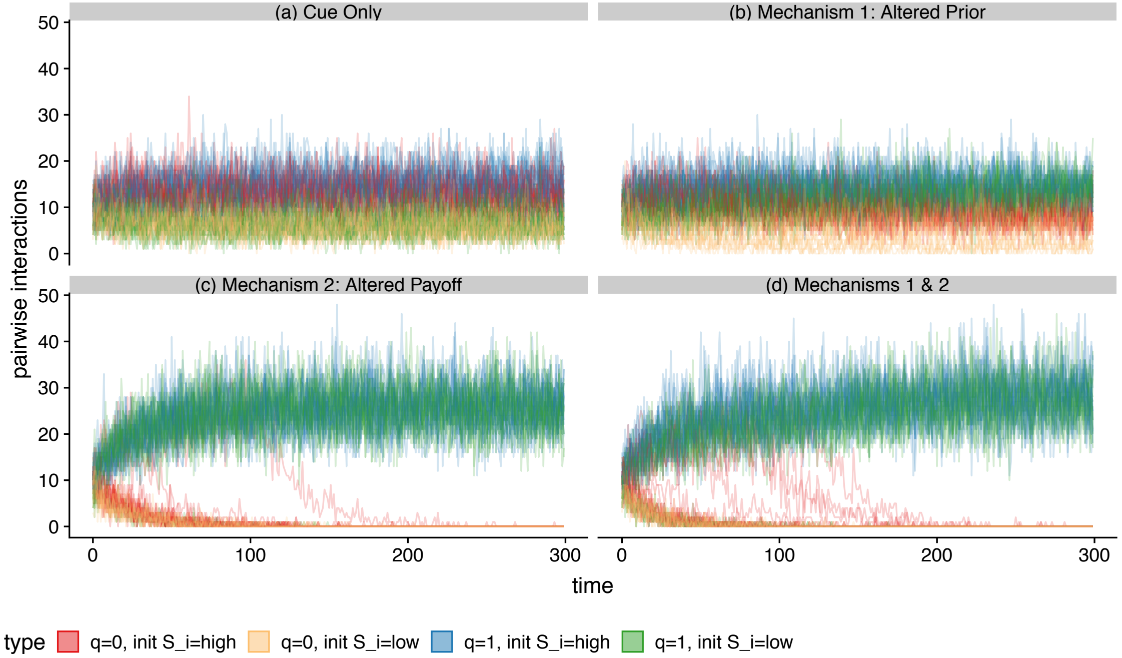
(a) Representative individual time series of S_i, for individuals of different quality and initial social prominence/capital.

(b) The distribution of S_i in each scenario, as a function of the individual's quality and initial social prominence/capital.

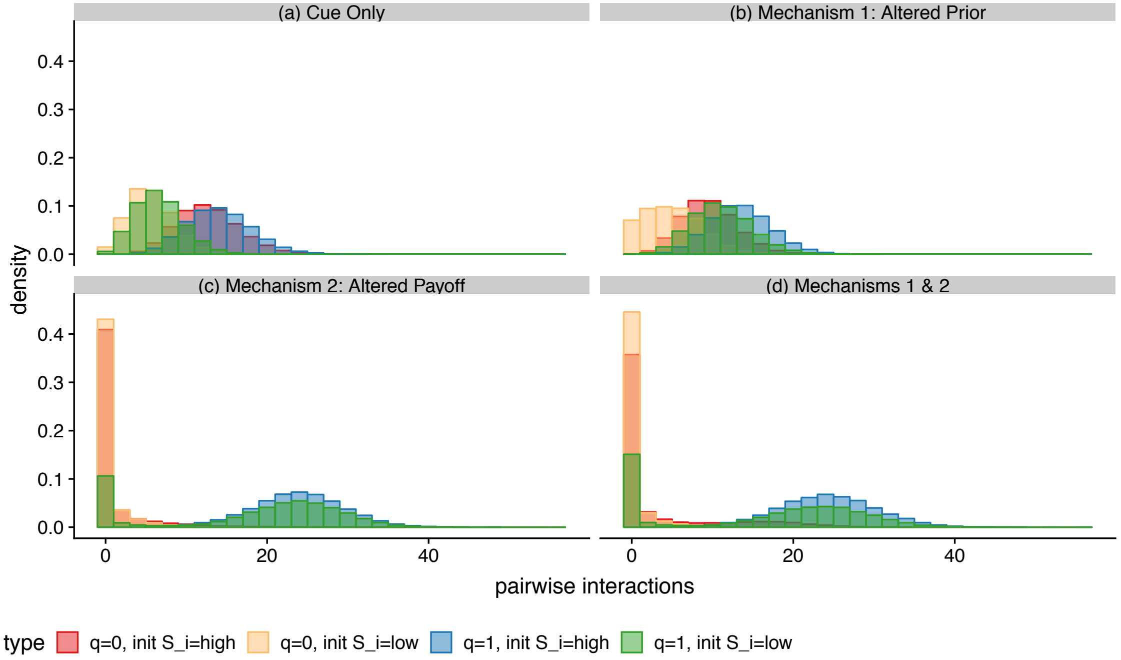
Figure SI.4: Agent based model results, for alternative scenarios

Signalling with both the altered prior and altered payoff mechanisms in combination in the absence of private learning further contributes to the "reputational shield." This is because the high social prominence/capital low-quality individuals do not get revealed during pairwise interactions.

In the main text, we indicated that the altered payoff mechanism increases the inequality in social prominence/capital via a feedback happening through pairwise interactions. This is illustrated in Figure SI.5. It shows that high quality individuals who start with high social prominence/capital accrue more pairwise interactions than they do in the other scenarios. This is due to the fact that other high quality individuals are stuck in a reputational poverty trap, so they get very low numbers of interactions, which accrue instead to high social prominence/capital individuals. This leads to substantial inequality in the number of interactions that individuals have.



(a) Representative individual time series of the count of pairwise interactions, for individuals of different quality and initial social prominence/capital.



(b) The distribution of the number of pairwise interactions in each model, as a function of the individual's quality and initial social prominence/capital.

Figure SI.5: Distribution and changes in the number of pairwise interactions for the scenarios in Figures 3 & 4 of the main text.

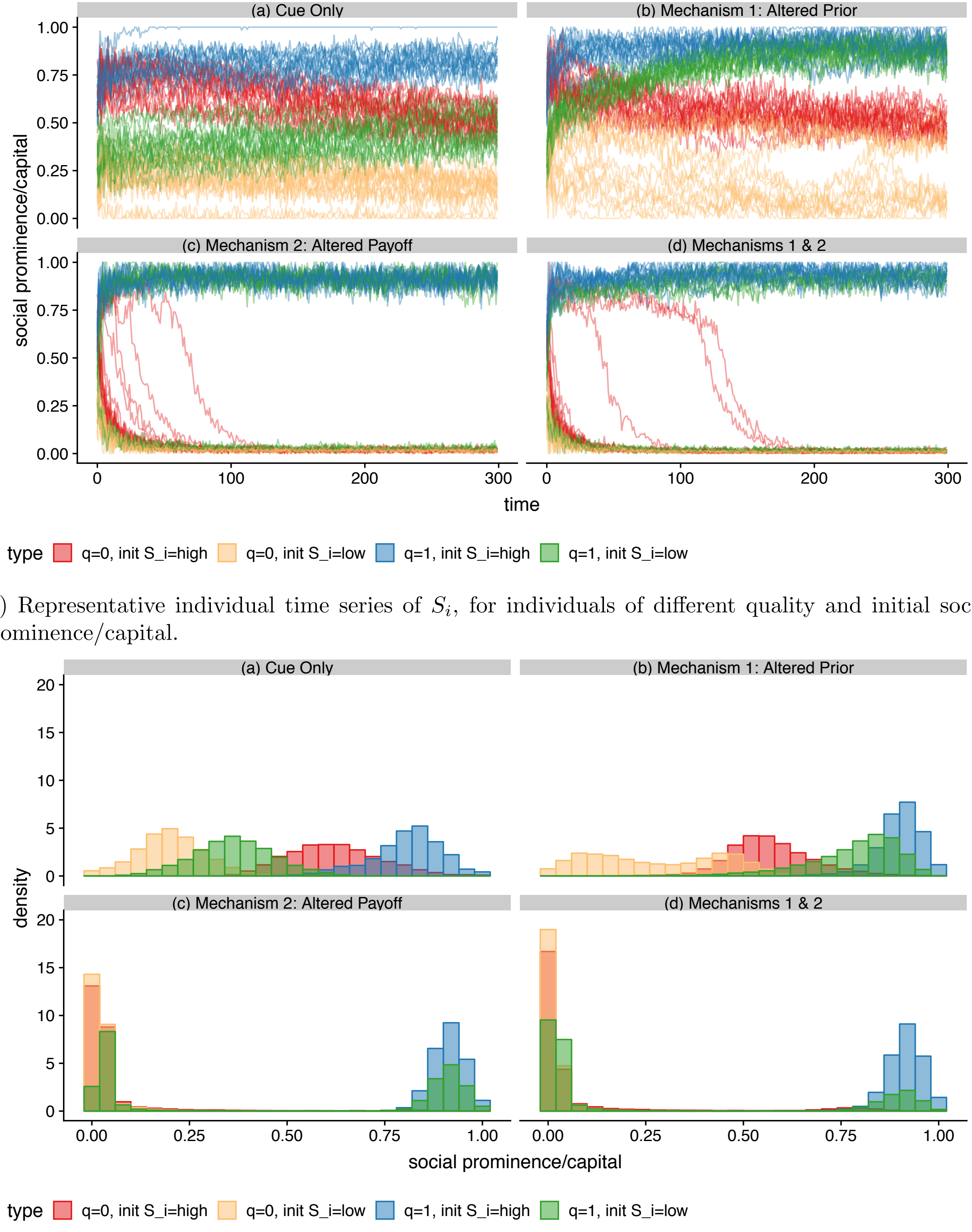
2.3 Robustness of results

Here we investigate how our conclusions change as we modify some aspects of the model.

2.3.1 Mutual learning in the pairwise interactions

The private learning during the pairwise interaction is one-sided in the main model: if i initiates an interaction with j, then only i learns about the quality of j (we can think of i making a request to j and learning how cooperative or resourceful j is). Now, instead, we can assume that both individuals can learn about each other. Figure SI.6 below shows the outcome of this alternative modelling assumption. Here, we allow both i and j to learn about each other when i initiates an interaction with j (through the same inference procedure as described above).

We see that this assumption does not change the fundamental dynamics. The main difference is that private learning plays a slightly more important role with this alternative. In the cue only model, there is a little bit faster learning about the real quality of individuals whose initial social prominence/capital hides their real quality. Similarly, the "fall from grace" is faster for low quality/high initial social prominence/capital individuals in scenarios that include the altered payoff mechanism. And, when the altered payoff mechanism is in play, more high quality/low initial social prominence/capital individuals fall into the reputational poverty trap. While the mutually informative pairwise interactions means there is slightly more movement in S_i for low quality individuals (and high quality individuals caught in the trap), it is still slight and insufficient for them to break out.



(a) Representative individual time series of S_i, for individuals of different quality and initial social prominence/capital.

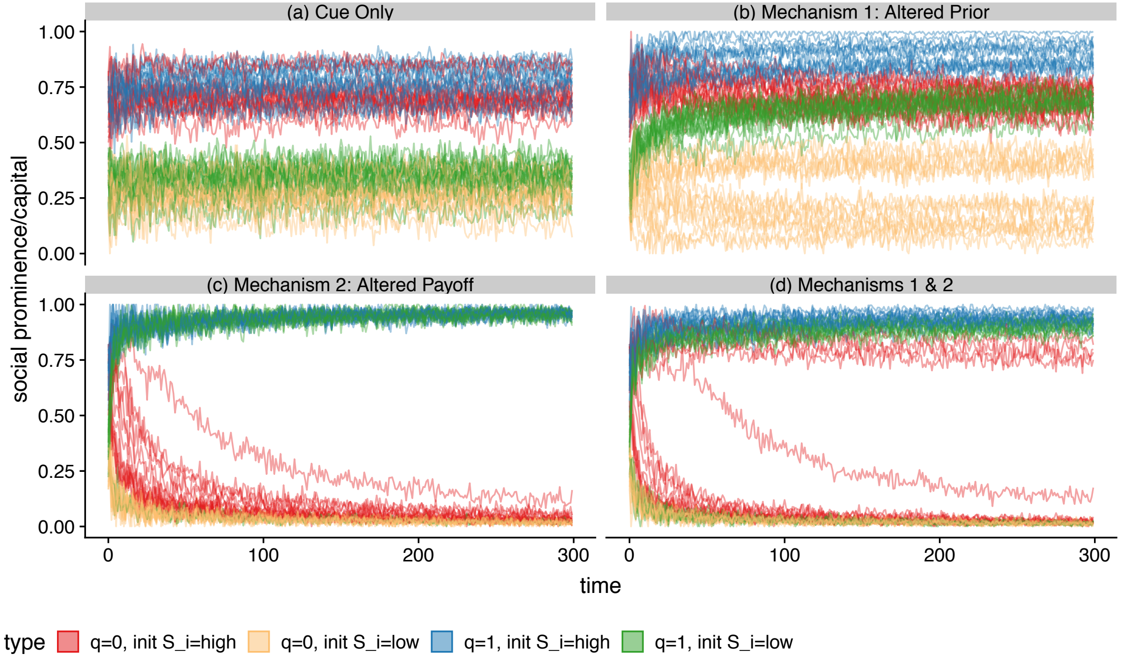
(b) The distribution of S_i in each model, as a function of the individual's quality and initial social prominence/capital.

Figure SI.6: Agent based model results when the pairwise interactions allow both parties to learn about each other, for the same set of scenarios as in Figures 3 & 4 of the main text.

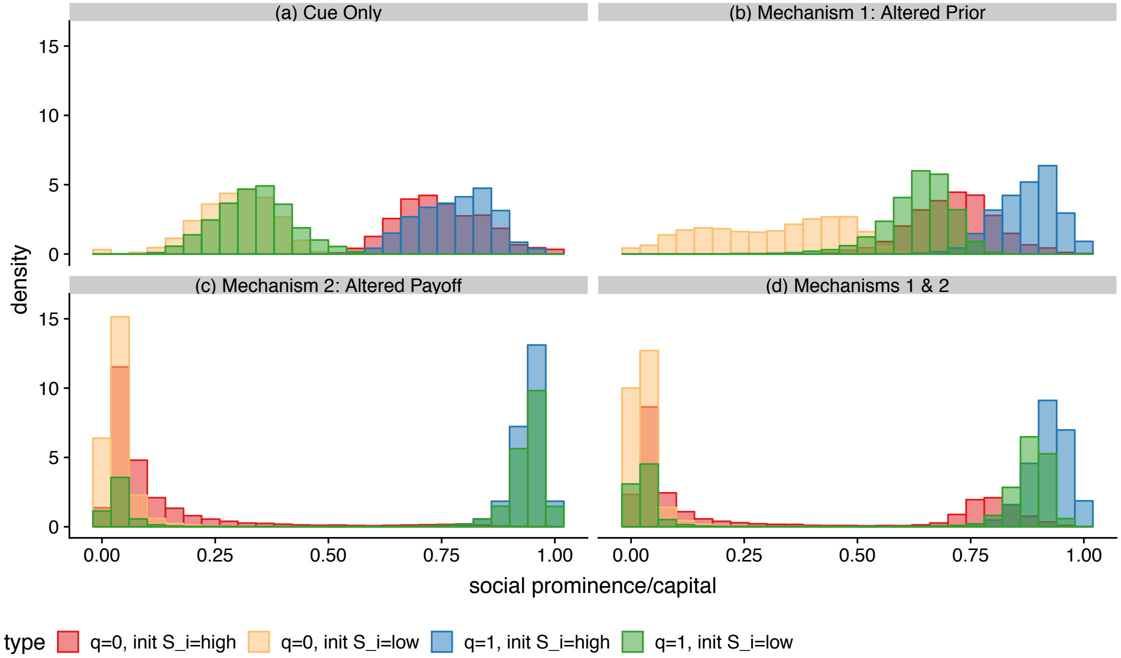
2.3.2 Effect of memory

When memory is perfect, initial impressions linger and so generally slow down the process of revelation. The lingering effect of low initial social prominence/capital is seen most clearly with the altered prior mechanism, with those individuals with low quality and high initial social prominence/capital retaining their high social prominence/capital over time, and individuals with high quality/low initial social prominence/capital increasing their prominence very gradually relative to the case where $\delta < 1$.

When the altered payoff mechanism operates, perfect memory generally slows the "fall from grace" for low quality/high initial social prominence/capital individuals. Most notably, when the two mechanisms work in tandem, these individuals can end up with almost any possible value of S_i: while some bottom out, others are able to maintain the reputational shield.



(a) Representative individual time series of S_i, for individuals of different quality and initial social prominence/capital.

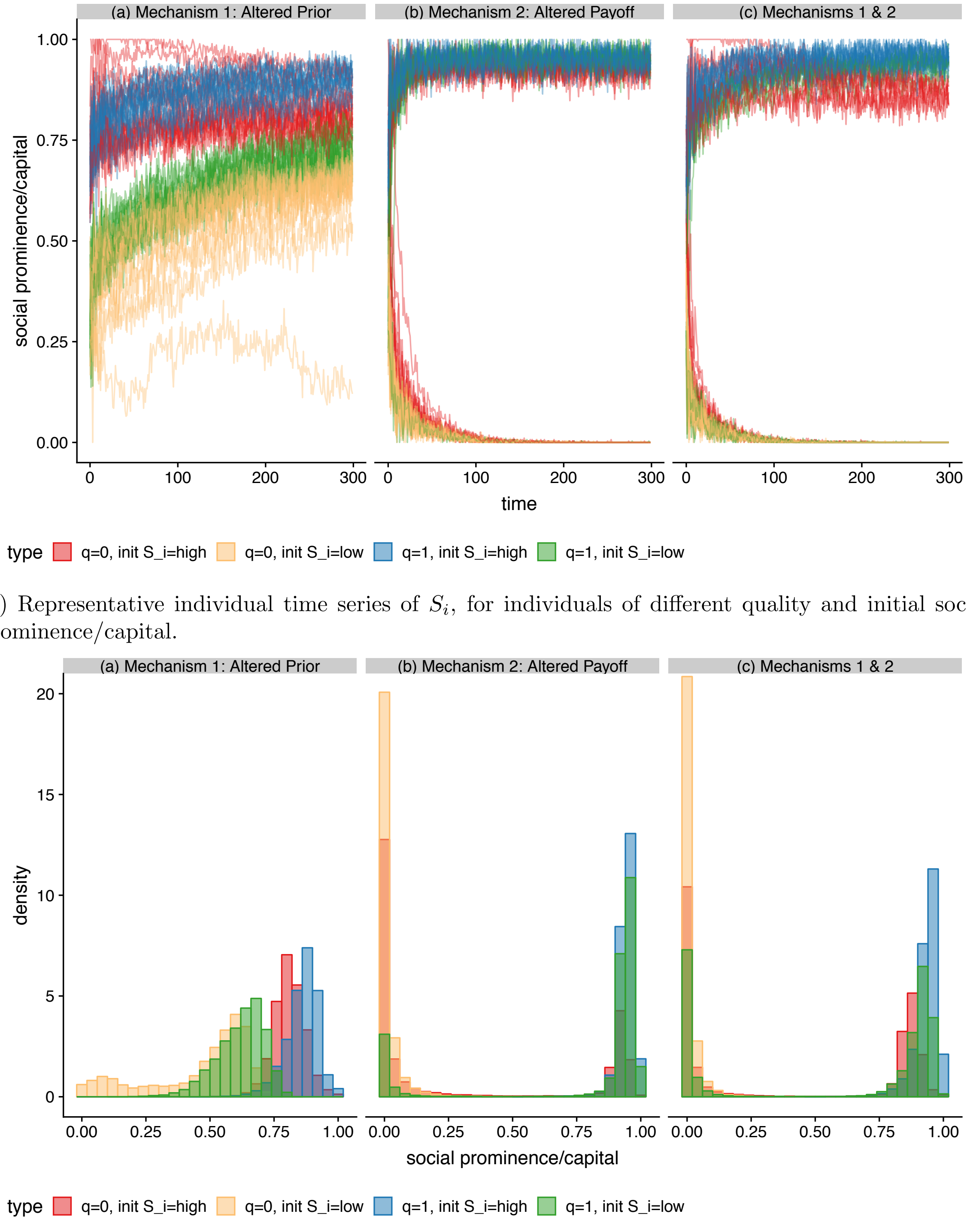


(b) The distribution of S_i in each model, as a function of the individual's quality and initial social prominence/capital.

Figure SI.7: Agent based model results, with perfect memory, for the same set of scenarios as in Figures 3 & 4 of the main text.

2.3.3 Success and failure unrelated to quality

Here we present the results of model runs where the probability of success and failure is unrelated to quality. When this is the case, the quality of individuals whose S_i is high enough to induce them to play pooling equilibrium strategies is no longer distinguishable by observers. The consequence is that the "reputational shield" is very robust (unless we increase greatly the number of bilateral interactions and reduce the noise in the learning during those bilateral interactions). As we see under the altered prior mechanism, individuals who have an S_i high enough to play pooling equilibrium strategies seem to converge towards a similar distribution of S_i independent of quality.



(a) Representative individual time series of S_i, for individuals of different quality and initial social prominence/capital.

(b) The distribution of S_i in each model, as a function of the individual's quality and initial social prominence/capital.

Figure SI.8: Agent based model results, when success and failure are not related to quality $\theta_H = \theta_L$, , for the same set of scenarios as in Figures 3 & 4 of the main text.

3 Statement on Citation Diversity

Recent work in several fields of science has identified a bias in citation practices such that papers from women and other minority scholars are under-cited relative to the number of such papers in the field (Mitchell et al., 2013; Dion et al., 2018; Caplar et al., 2017; Maliniak et al., 2013; Dworkin et al., 2020). Here we sought to proactively consider choosing references that reflect the diversity of the field in thought, form of contribution, gender, race, ethnicity, and other factors. First, we obtained the predicted gender of the first and last author of each reference by using databases that store the probability of a first name being carried by a woman (Dworkin et al., 2020; Zhou et al., 2020). By this measure (and excluding self-citations to the first and last authors of our current paper), our references contain 13.59% woman(first)/woman(last), 12.94% man/woman, 3.95% woman/man, and 69.52% man/man. This method is limited in that a) names, pronouns, and social media profiles used to construct the databases may not, in every case, be indicative of gender identity and b) it cannot account for intersex, non-binary, or transgender people. Second, we obtained predicted racial/ethnic category of the first and last author of each reference by databases that store the probability of a first and last name being carried by an author of color (Ambekar et al., 2009; Sood and Laohaprapanon, 2018). By this measure (and excluding self-citations), our references contain 7.94% author of color (first)/author of color (last), 14.21% white author/author of color, 15.04% author of color/white author, and 62.8% white author/white author. This method is limited in that a) names and Florida Voter Data to make the predictions may not be indicative of racial/ethnic identity, and b) it cannot account for Indigenous and mixed-race authors, or those who may face differential biases due to the ambiguous racialization or ethnicization of their names. We look forward to future work that could help us to better understand how to support equitable practices in science.

References

Ambekar, A., Ward, C., Mohammed, J., Male, S., and Skiena, S. (2009). Name-ethnicity classification from open sources. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 49–58, New York, NY, USA. Association for Computing Machinery.

Caplar, N., Tacchella, S., and Birrer, S. (2017). Quantitative evaluation of gender bias in astronomical publications from citation counts. *Nature Astronomy*, 1(6):1–5.

Dion, M. L., Sumner, J. L., and Mitchell, S. M. (2018). Gendered citation patterns across political science and social science methodology fields. *Political Analysis*, 26(3):312–327.

Dworkin, J. D., Linn, K. A., Teich, E. G., Zurn, P., Shinohara, R. T., and Bassett, D. S. (2020). The extent and drivers of gender imbalance in neuroscience reference lists. *Nature Neuroscience*, 23(8):918–926.

Maliniak, D., Powers, R., and Walter, B. F. (2013). The gender citation gap in international relations. *International Organization*, 67(4):889–922.

Mitchell, S. M., Lange, S., and Brus, H. (2013). Gendered citation patterns in international relations journals. *International Studies Perspectives*, 14(4):485–492.

Sood, G. and Laohaprapanon, S. (2018). Predicting race and ethnicity from the sequence of characters in a name. *arXiv:1805.02109 [stat]*. arXiv: 1805.02109.

Zhou, D., Cornblath, E. J., Stiso, J., Teich, E. G., Dworkin, J. D., Blevins, A. S., and Bassett, D. S. (2020). Gender diversity statement and code notebook v1.0.