CrossMark

Honor and Violence

An Account of Feuds, Duels, and Honor Killings

John Thrasher [1] ⓘ · Toby Handfield [2]

Abstract

We present a theory of honor violence as a form of costly signaling. Two types of honor violence are identified: revenge and purification. Both types are amenable to a signaling analysis whereby the violent behavior is a signal that can be used by out-groups to draw inferences about the nature of the signaling group, thereby helping to solve perennial problems of social cooperation: deterrence and assurance. The analysis shows that apparently gratuitous acts of violence can be part of a system of norms that are Pareto superior to alternatives without such signals. For societies that lack mechanisms of governance to deter aggression or to enforce contracts, norms of honor can be a rational means of achieving these functions. The theory also suggests that cultures can become trapped in inefficient equilibria owing to path-dependent phenomena. In other words, costly signals of honor may continue to be sent even when they are no longer providing useful information.

Keywords Honor · Violence · Honor killing · Social norms · Signaling

The idea of "honor" is a diverse notion. We speak of "honoring a debt," or of "paying someone an honor" when speaking highly of them. We award honorary titles and positions to those we wish to publicly applaud. We justify sticking to our principles, even at personal cost, by saying it is a "matter of honor" or that it is required by our "code of honor."

✉ John Thrasher
Johnthrasher23@gmail.com

[1] Philosophy Department & The Smith Institute for Political Economy and Philosophy, Chapman University, Becket Building, One University Drive, Orange, CA 92866, USA

[2] Philosophy Department, Monash University, 20 Chancellors Walk, Menzies Building, Clayton, VIC 3800, Australia

Springer

Honor is also implicated in a range of violent behavior: what we call *honor-based violence*. This violence comes in many forms throughout history and across the globe. Feuds, vendettas, and duels are classic examples. Another form of honor-based violence is the murder of women who have violated or are believed to have violated local norms of sexual behavior: "honor killings." Honor violence is interesting because it is ubiquitous (occurring throughout history and in many different parts of the world), similar (despite local differences, many core elements of honor-based violence are similar from culture to culture), and seemingly gratuitous (the amount of violence involved seems difficult to justify based on any conceivable benefits).

No coherent and complete theory of honor has been proposed to explain both honor-based violence and the concepts denoted by the more common uses of the word "honor." Philosophers have approached the topic by attempting to distinguish honor from closely related concepts of esteem and respect (Appiah 2011; Sessions 2010) and undertaking normative evaluation of honor, relating it to other normative systems such as morality and religious doctrine (Appiah 2011; Demetriou 2014; Doris and Plakias 2008; Kumar and Campbell 2016). Social scientists, following Nisbett and Cohen's (1996) influential work, have attempted to identify and quantify the psychological mechanisms that mediate honor-related behavior (Benard 2013, 2015; Grosjean 2014; Shafa et al. 2014). While both approaches are illuminating, neither the conceptual nor the psychological approach to honor provides an ultimate explanation for why honor cultures exist in the first place. The most promising work in this vein to date investigates the norms and behavior of local subcultures that embrace norms of honor despite the surrounding society having relatively little role for them. These subcultures include professional criminals, prisoners, and terrorists (Chen 2010; Gambetta 2009; Iannaccone and Berman 2006; Skarbek 2011, 2014). Case studies such as these provide a natural experiment of sorts: they enable the development of models that might explain the ultimate causes of honor-based violence. We call this the *comparative-functional* approach to the study of organized honor violence.

Drawing on models and insight from recent work on violence, we develop a unified account of honor norms that incorporates all three of these approaches. Honor norms, we argue, are a class of social norms that perform important governance functions in societies with weak mechanisms for organizing and controlling endogenous violence. *Honor-based violence* is a signal (not always truthful) of quality or status and *norms of honor* are social mechanisms for both reliably creating and communicating that signal and also reinforcing the social structure of status that the signals create.

Our approach can explain the ubiquity and similarity of honor-based violence across time, space, and culture. Honor norms are ubiquitous because they are a stable solution to common human problems in environments without effective exogenous governance mechanisms. These norms are not the same everywhere, and they do not always coexist, but they tend to share certain reliable characteristics. In particular, we argue that honor norms can be divided into two general types—*revenge* and *purification* norms—based on the functional role that each plays. We see these two general types of norms in many different cultures and at different times in history. In this paper, we provide a theoretical framework and model that can be used by researchers in various fields to expand our understanding of honor-based violence in its diverse incarnations.

Honor Norms

In our account, honor norms organize cooperation and conflict by reference to socially recognized measures of quality. What "quality" amounts to will vary depending on the social context and the particular honor code and norms of a particular society. For women, quality is, in many societies, related to their sexual fidelity, which in turn reflects their value on the marriage market or as wives. For men, it is often based on their trustworthiness to abide by and defend the norms of honor that require risk and danger—the ability to reliably deploy violence only when the norms require.[1] As we will argue below, honor norms can be divided into two types, each of which uses a reputation or signaling mechanism to communicate information about different types of quality.

Honor-based violence norms can be divided in two general types: *revenge* and *purification*. The first involves defending one's honor by assaulting or killing someone who has insulted or aggressed against oneself or one's close associates. More precisely, revenge-type violence involves violence that is instigated by an insult or act of an out-group member that lowers one's social status. Revenge-type violence is intended to regain that lost status and to restore one's "honor" in the eyes of the relevant members of the in-group and out-group. This sort of honor violence is commonly associated with feuds or vendettas, which entail cycles of retaliation and counter-retaliation. Engaging in retaliatory violence is necessary if one's family, clan, or gang is to avoid a significant loss of standing: a stain on its honor. These phenomena have been documented at various historical times among a variety of peoples, such as Pashtun (Ginsburg 2011), Turkana (Moritz 2008), Montenegrins (Boehm 1986), white males of the US South (Nisbett and Cohen 1996), a variety of ethnic groups around the Mediterranean (Black-Michaud 1975; Schneider 1971), and criminal gangs (Skarbek 2014).

The second type of honor violence is honor through *purification*. It typically involves the beating, disfiguration, or killing of a member of one's group. As with revenge-type violence, purification-type violence is triggered by an act that will result in a loss of social status (i.e., honor). Unlike revenge-type honor, however, the act in question is committed or believed to have been committed by an in-group member and the retaliatory violence is also directed at an in-group member. Where the family is the most important group, violence will usually be directed at a close relative—almost always female—for having transgressed strict norms of sexual chastity. These "honor crimes," unlike the acts of violence that are associated with revenge norms, target victims within the perpetrator's most intimate in-group. Killing the alleged transgressor is thought to be essential to maintaining the honorable standing of the family, and to removing the extreme shame that is attached to the transgression. Honor killings have been documented most frequently in Middle Eastern and South Asian countries, predominantly but not exclusively in Muslim communities (Chesler and Bloom 2012), and have also occurred among diaspora of these communities in Europe, North America, and elsewhere (United Nations Population Fund 2000).

---

[1] In honor societies, mad dogs are as dangerous as cowards, if not more so. Many of the norms regulating prison gangs, for instance, are related to preventing unauthorized violence between and within the different gangs. As David Skarbek (2014) shows in great detail in his study of the norms of American prison gangs, rules regulating violence are often highly specific about who may use violence on whom and under what circumstances. As Skarbek (2014:86) notes, in prisons, the gang "must authorize the use of violence because spontaneous, unplanned violence causes problems for other inmates."

In prison gangs, purification norms and rules are used to police reputation and to stabilize interaction and trade between groups. Purification violence, in this context, is a form of in-group policing that can be remarkably effective in the presence of noisy signals or unreliable information (Fearon and Laitin 1996). Prison gangs operate on a "community responsibility" system whereby "Each gang is responsible for their members' actions, so they have an incentive to monitor their members to ensure they maintain their collective reputation" (Skarbek 2014:83). We find the same type of mechanism of collective responsibility and reputation used, for similar reasons, by medieval merchants (Greif 2006:96–97). The punishment of plagiarists in an academic setting also works this way and is frequently prescribed by what is called an "academic honor code." Professional standards boards frequently inflict sanctions—the most severe being expulsion—on members who have behaved in ways that damage the reputation of the profession.

The key taxonomic difference between our two paradigms is the site of the sanction. *Revenge*-type honor norms require sanctioning an out-group member, and the ostensible purpose of the violence is to revenge an earlier wrong. *Purification*-type honor norms require victimizing an in-group member, and the ostensible purpose is a type of penance or purification. In the next two sections, we elaborate the functional differences between these two types of honor violence, and the way they are related to signaling systems.

Revenge-Type Honor Norms—The Deterrence Problem

Revenge-type honor norms address one of the oldest problems of human society: how to avoid the Hobbesian generalized war of all against all (Hobbes 2012). This problem—the *deterrence problem*—requires establishing a credible threat that violations of one's self or property will be met with sufficient violence so as to deter first strikes. Because retaliation is costly, any effective deterrence must convince prospective aggressors that an attack will be met with retaliation, despite the costs. To do this, the potential victims of a first strike need to signal that they are not "rational" in this sense: they are willing to fight even when the cost of fighting is higher than the value of the good to be defended.

A Model of Revenge-Type Violence

One well-known model that explains the rationality of this seemingly irrational commitment to deterrence was proposed initially to address a related paradox in commerce: why a firm would ever engage in a price war so as to deter competition (Kreps and Wilson 1982; see also Selten 1978 for the original puzzle). The game is asymmetric, involving two roles: challenger and incumbent (Fig. 1). The challenger has an opportunity to make some incursion against the territory or rights of the incumbent. If no challenge is made, the status quo is preserved. If a challenge is made, the incumbent must decide whether to retaliate or submit. Retaliation is costly, but it also deprives the challenger of material gain from the interaction.

Now further suppose that there is a plurality of incumbent types. This introduces an additional asymmetry. The typical incumbent is *weak* and has the payoffs that favor
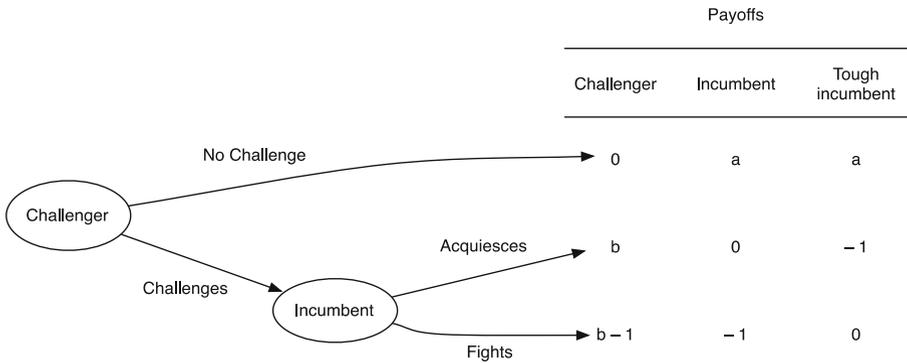
Fig. 1 Payoffs for challenger and two types of incumbent in Kreps and Wilson's game. $0 < a,\ b < 1$

acquiescence. With some small probability, unobserved by the challenger, an incumbent may be *tough*. The tough incumbent expects to benefit from violent conflict because of its strength, or at least finds it rewarding to retaliate because of its psychological constitution.

Given these payoffs, a tough incumbent should retaliate. And if the encounter was a one-off affair, then weak incumbents should certainly capitulate. In a setting where the encounters will be repeated, however, if a weak incumbent retaliates, it may leave the challenger uncertain as to what type the incumbent is. By contrast, *failing* to retaliate will leave the offender in no doubt that the incumbent is weak. Hence after one capitulation, a weak incumbent can expect to suffer challenges on every future opportunity.

The parallel to honor is clear. Failing to retaliate entails a loss of social status: a loss of honor. Costly violence is necessary to preserve the uncertainty others have about the incumbent's willingness to retaliate in future. In effect, in this model, being honorable is simply the status of not (yet) having been proven weak. This model explains a number of features of revenge-type violence.

(1)   Any family that has suffered a transgression risks a loss of honorable status unless it retaliates. But if the act of retaliating in turn jeopardizes the honor of the challenger, then we see the origins of the turn-based structure of strike and counterstrike, typical of feuds and vendettas. If the violence was in direct pursuit of territory or other resources, then we would expect it either to escalate to full-fledged war or to diminish to naught. But because the violence is a matter of reputation repair, only one family needs to prove its honorable status at any given time, so only one family has an incentive to carry out the next act of aggression. Hence the violence is protracted, but not necessarily escalated. Even in the absence of a feud, some ethnographers report a relentless struggle for maintenance of honor and vigilance against threats:

The game was a laborious one because it demanded the greatest sensitivity to insult and challenge and because there were no intermissions once it started in earnest at the onset of physical maturity. . . . The interminability of the enterprise is but one reason why this [word] "game" needs [to be in] quotes. It was a game only in the sense that honor necessarily meant competition. There was nothing

trivial about the "game"; it was, for people of self-respect, coterminous with social existence itself (Miller 1990:31).

(2)   The Kreps and Wilson model predicts that even weak incumbents will engage in a lot of retaliation before finally capitulating. Only in the end-game, where few future challenges are anticipated, will weak and tough types have sufficiently different incentives that their behavior diverges. This explains why we expect nearly everyone in honor societies to be "honorable." The few families who are not able to maintain the required violence are shamed, ostracized, and eventually leave (Mahdi 1986). But this does not mean all the remaining families who are still in good standing are in fact "tough."

Diego Gambetta describes this process of filtering among prisoners, especially new prisoners, who must immediately show other prisoners that they have the courage and resolve to defend themselves. According to Gambetta and the prisoners he quotes, new inmates are immediately tested with respect to their propensity to retaliate. Relatively small violations or assaults are used as a litmus test for the quality of the prisoner in the deterrent sense (Gambetta 2009:99). As Gambetta argues, these tests create a hierarchy within the prison that acts to "sort prisoners initially into two main types—not so much winners and losers, but fighters and passive victims" (2009:100).

(3)   A number of writers have noted that the strategic interactions that affect honor appear to be zero-sum or even "negative sum." This accords with the core idea of the Kreps and Wilson model, that violent retaliation is explained by a fundamentally competitive setting.

The amount of honor in the Icelandic universe was perceived to be constant at best, and over the long run, it seemed to be diminishing. The men of the present generation were never quite the men of their great-grandfather's time (Miller 1990:31).

If honor is merely the status of not yet having been proven to be weak, then it is not surprising that its sum appears to diminish within just a few generations, as individuals fall into shame, but none (or very few) are promoted from shame to honor. With respect to the intergenerational perspective reported by Miller, this may be because, unlike for the living, there is no fierce demand to identify the weaknesses of the dead. Hobbes (2012:152, chapter XI) goes further, suggesting that there may be a strategic advantage in valorizing the dead in a competitive struggle for social status such as honor:

Particularly, competition of praise, enclineth to a reverence of Antiquity. For men contend with the living, not with the dead; to these ascribing more than due, that they may obscure the glory of the other.

Norms and Reinforcement

We conjecture that the distinctive *norms* of honor reinforce the above strategic incentives, especially when collective defense of honor is important. Social norms are shared preferences for behavior that are conditional on others having similar preferences (Bicchieri 2006, 2016) and can be powerful tools for overcoming social dilemmas (Bicchieri 2002). Agents may want to behave fairly in a public goods game, for instance, not because they are intrinsically motivated to be fair but because they believe everyone else expects them to behave fairly. Collectives that share a common norm of this sort can coordinate on socially optimal outcomes, discouraging free riders because of the conditional nature of their commitment.

As already noted, a group that wishes to solve the deterrence problem faces a collective action problem and this same problem arises at the moment of signaling: all members of the collective benefit from the signal, but the individual who takes provocative or risky action bears a disproportionate part of the cost. Norms of honor motivate individuals to take on these costs: e.g., I prefer to defend the honor of the family, conditional on enough of my fellow family members defending our honor. Transgression from these norms brings intense social sanction. Ethnographic evidence supports the notion that there is significant social pressure on individuals to defend family honor (Boehm 1986:58–63) and that the decisions to undertake violence are not always taken in hotheaded fury, but sometimes reluctantly, in response to anticipated shame and disapproval.

An example of this can be seen in the ostracism practices of the Pathan tribes in northwest Afghanistan, who rely on an honor code (Pukhtunwali) to govern their interactions.

> To act contrary to Pukhtunwali is to be dishonorable, and with that stigma, it becomes virtually impossible to function in tribal society. . . . Ostracism, as a response to the contravention of Pukhtunwali (or any established local custom), becomes the obligation of every Pathan, acting individually or as part of a relevant tribal segment. In carrying out the dictates of Pukhtunwali, a Pathan consciously may be discharging his obligation only for the sake of personal izzat [honor], but in effect he is acting as an agent of the tribe to control the aberrant behavior that threatens the viability of the tribal structure (Mahdi 1986:297).

This combination of strategic incentives to use violence as a signal as well as the social norm to motivate individuals to defend collective honor is mutually reinforcing. The norm explains how a dangerous family overcomes the motivational problem to send a costly signal. But if the norm is established, it may explain part of what makes a family a dangerous subject of transgression in the first place, creating a positive feedback loop (Fig. 2). This feedback loop may explain why we often see an apparently extreme and rigid commitment to retaliatory violence in honor cultures.

Without supposing that it is the *best* solution to the Hobbesian problem of deterrence (the property-rights apparatus of a modern nation-state is almost certainly superior), we should be open to the possibility that honor violence is not entirely gratuitous (cf. Brennan et al. 2013:143).
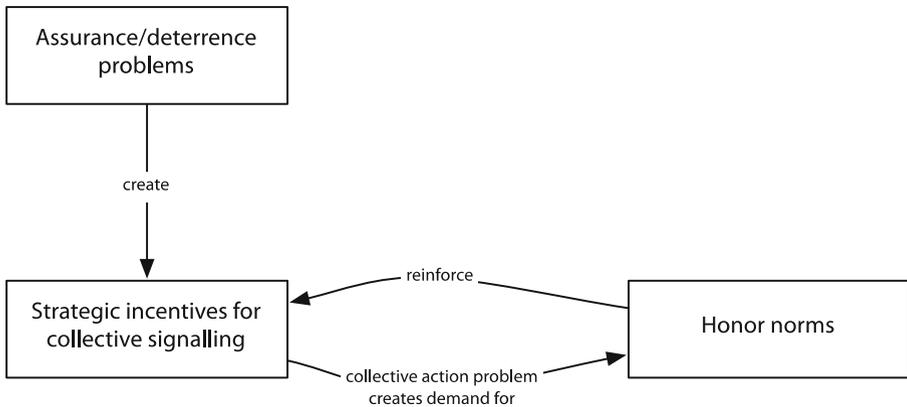
Fig. 2 The explanatory relations between the deterrence/assurance problems, strategic incentives to signal, and norms of an honor culture

Purification-Type Honor Norms—The Problem of Assurance

Purification honor violence is directed at in-group members who are alleged transgressors of an honor norm. One obvious hypothesis is that it functions as a means of *punishing* the transgressor, thereby deterring similar transgressions in the future, both from the specific individual who transgressed and from other group members who might consider similar behavior. This deterrence account of purification violence is surely part of the story, linking it with revenge-type norms, but it is incomplete. First, the deterrence account struggles to explain the extremely high cost involved in some purification violence—which includes lethal sanctions in response to mere sexual transgressions. Although many societies have codes of sexual behavior, it is relatively rare that they resort to such extreme measures to enforce them. It is implausible that the temptation to sexual deviance is inherently greater in these societies. It is no doubt true that the incentive to demand compliance with sexual codes is often very high, but that merely re-describes the phenomenon rather than explaining it. We ultimately want to know why there are such relatively strict norms regulating sexual behavior in the first place. Second, if purification violence is explained by a deterrence function, it would suggest that the violence needs to be witnessed only by other members of the in-group (e.g., family members). But examination of reported cases shows that, overwhelmingly, these murders are undertaken with the intention of achieving some degree of publicity in the immediate community (Kulczycki and Windle 2011).

We hypothesize that, in addition to any punitive role it may have, purification violence possesses a *signaling function*, directed at the relevant community, of the family's quality as a future partner in important economic and social transactions—particularly marriage, in the case of honor killings. This type of norm is not limited to honor killings, however. Indeed, it is helpful to look at an apparently benign institutional example that has a structurally similar dynamic: the expulsion of professionals for misbehavior by their relevant professional societies. Having identified the way in which purification sanctions can signal in that setting, it will be easier to see parallels in cases of outright violence.

Purification-type honor norms involve a *signaling* mechanism.[2] Signaling involves a sender successfully communicating information to a receiver (Skyrms 2010). Suppose that there are two types of sender, high and low quality, such that receivers prefer to interact with high-quality types, and high-quality types stand to gain more from successful interaction with a receiver than do the low-quality types: $b_H > b_L$. High-quality types wish to send a credible signal to receivers indicating that they are of high quality, but low-quality types also have an incentive to masquerade as high-quality types. If receivers only respond to a signal that costs some amount c, which is greater than b_L, then it will not be worthwhile for the low-quality types to attempt to deceive. Further, if the cost of the signal is lower than b_H, it will still be worthwhile for the high-quality types to send the signal. So if $b_H > c > b_L$, signaling can be a stable equilibrium. Only high-quality types signal, and receivers always believe that the signal is indicative of high quality.

The above describes a simple signaling game that involves differential benefit for high- and low-quality types (Johnstone 1997; Getty 1998).[3] Our contention is that purification-type honor violence serves as a costly signal of a sort of "quality" that is of great importance in the economic and social circumstances of an honor society.

Suppose there are two types of professionals competing in a market. The high-quality type makes fewer errors, and customers therefore prefer to match with them. But customers cannot easily observe quality, so they rely on reputational information to make their decision. This motivates professionals to form voluntary associations of individuals of roughly commensurate quality that can aggregate their reputations and provide a reliable source of monitoring.[4] When a transgression occurs (negligence, corruption, misconduct), the professional body is motivated to respond in a way that preserves the reputation (and by extension the market share) of its members at the highest possible level.[5] Expelling a member may, from this point of view, seem a poor choice of strategy. Decisions of this sort are likely to be legally disputed and to attract significant publicity. Simply tolerating the transgression may lead to many fewer individuals knowing about the problem and be a tempting alternative. But beyond some threshold level of publicity for a transgression, the association may hope to gain more from publicly sanctioning the responsible individual than from further attempts to conceal, justify, or otherwise minimize the transgression.

If high-quality groups have this sort of incentive to undertake costly punishment, however, it may be tempting for a lower-quality group to emulate their behavior. If the signal sent by punishing can be faked by low-quality associations, it will be much less attractive for a high-quality organization to employ this strategy. So, the question arises whether there is a way of credibly signaling high quality by one's choice of sanctioning strategy.

---

[2] In some sense, the model we used to illuminate revenge-type violence involves signaling also, but game-theorists refer to such phenomena as involving *reputation*, rather than signaling: the "message" is intrinsic to the very behavior.

[3] Other signaling equilibria are possible in situations in which both high- and low-quality types stand to make identical gains, but high-quality types can signal at lower *marginal* cost (Grafen 1990; Spence 1973) or even at zero cost (Gintis et al. 2001; Lachmann et al. 2001).

[4] Such organizations also provide opportunities for gratifying mutual recognition and conferral of esteem (Brennan and Pettit 2006, chap. 11 §1).

[5] This is another example of the "community responsibility" system of reputation management that we see in prison gangs and medieval traders (Greif 2006; Skarbek 2014:75–77).

If both low- and high-quality groups stand to make similar gains from signaling, and will pay similar costs, then theory predicts that we are unlikely to see a separating equilibrium: we instead expect to see pooling equilibria in which every group signals or none do. But a difference in expected benefit between the groups is plausible because a higher-quality group anticipates a relatively low rate of scandals in the future. Thus, the expected benefit of assuring receivers to continue cooperating is large. A lower-quality group runs a significant risk that any benefit achieved by signaling now will be undermined by a future misadventure. If the future misadventure goes unpunished, they will have reaped a smaller *benefit*. If future misadventures must be punished, the group is shouldering a significantly higher *cost* for maintaining commensurate sanctions with the high-quality group. So, a low-quality group either faces a higher marginal cost for maintaining a consistent punishment strategy or it has a lower expected benefit if it will later lose its reputation by failing to consistently punish subsequent transgressions. These circumstances suffice to make costly signaling rationally sustainable: it can be to the advantage of high-quality groups to undertake costlier sanctions than low-quality groups, and consumers thereby obtain some evidence of quality.

In societies where honor killings are common, marriages are a very significant social and economic transaction between families (Kulczycki and Windle 2011). And marriage contracts are a tricky business; each family has a limited number of attempts to broker marriage. Divorce rights are limited, and if divorce does occur it comes with significant social stigma. Marital arrangements in honor societies also tend to be highly patrilineal; wives are economically dependent on the husband's family. Young women in these societies are, in effect, treated as an especially valuable commodity for forging alliances and advancing the family's social position (Ginsburg 2011). If a daughter is perceived as transgressive, this affects the social standing and economic well-being of her extended family (Kressel et al. 1981; Weiner 2013). In this environment, individuals have reason to care about the behavior of their cousins, nephews, and nieces (not just the prospective bride and groom) in a way that would be largely irrelevant to anyone planning a marriage in a society with different norms.

In the case of obtaining services from a doctor or a lawyer, relatively objective criteria exist to determine what constitutes good-quality professional services. In a marriage market, because one wants to have children who are themselves regarded as desirable by others in that market, the qualities that determine the desirability of a mate are socially constituted. Agents do not merely want what they regard as intrinsically desirable: they also want whatever they anticipate *others* will want. In such domains, certain types of social norm can play a very significant role in influencing behavior (Bicchieri 2016).

Gerry Mackie has argued that practices such as footbinding in China and female genital mutilation in Africa are extreme instances of the same dynamic: parents do not—at least initially—intrinsically value harming their children in these ways. Rather they believe that, given what everyone else prefers, it is essential to conform in order to secure their children's prospects of marriage (Mackie 1996).[6] These conventions can be shifted, but not readily. In the case of footbinding, it was through concerted efforts to shift society to a different convention that the practice diminished. But when it did so, it

---

[6] Mackie's account, while ground-breaking and highly influential, is not uncontroversial. See especially Efferson et al. (2015), Ross et al. (2016), and Howard and Gibson (2017) for recent work on this topic.

collapsed remarkably quickly (Mackie 1996:1001).[7] Societies that practice honor killing have a similarly strong preoccupation with a woman's sexual behavior. Women are frequently confined, veiled, and deprived of opportunities to work or study. A superior equilibrium is available in which these costly practices are not regarded as prerequisites for marriage but because any family that deviates from the conventions is likely to suffer immediate costs in the marriage market, it is difficult to shift.

With this notion of "quality" in play in the relevant marriage markets, a family in which one child has behaved dishonorably risks a loss of reputation that will be devastating to the marriage prospects of all the remaining children. A family that knows its other children have not committed any similar transgressions will wish to signal to potential spouses that they can be confident in the quality of the family. Killing a daughter who has transgressed is a horrific cost, but precisely because of its cost, it is a far more credible signal. A patriarch of a "low quality" family—one who knows that his other children will likely be discovered in future sexual transgressions—would not be able to "afford" such an extreme policy of sanction.

Again, it is possible that other mechanisms could be used to achieve this signal; all that is required, from a strategic point of view, is that the party who wishes to send the signal bears a cost. This explains the defining feature of violence related to purification-type norms: why it is inflicted on a fellow group member. In contrast to violence associated with revenge-type norms, there is a natural salience to signaling by way of punitive actions. First, it allows one costly act both to *deter* transgressions by in-group members and to *signal quality* to out-group members simultaneously, and these functions are likely to be mutually supporting. Second, there is a symbolic benefit in carrying out such an extreme sanction as murder: it categorically removes from the family the source of reputational threat.

Considered at a more abstract level, in both of these markets the central problem facing participants is one of assurance. There is an asymmetry of information in both cases. One side has better information than the other about the quality of what they have to offer, and compounding this, the side with less information is exposed to greater risk. The qualities that make a desirable husband in these communities (wealth, political power) are relatively easily observed and difficult to fake; the qualities that make a desirable wife (chastity, docility) are much harder to observe and easier to fake. Similarly, it is very difficult for a prospective client to observe the quality of a professional, and the client stands to lose much more than the professional by making the wrong choice. Credible signals of quality may arise in a number of ways in such markets.[8]

We have focused on a signal that can arise in the context of sanctioning transgressions. Even if both high- and low-quality groups stand to make equivalent gains from persuading others of their quality, high-quality groups will be able to sustain a higher

---

[7] Social norms, in this way, can often be fragile—when they change, they change quickly (Bicchieri 2006:181,197; Thrasher and Vallier 2015).

[8] Paralleling our model of honor killing, Rai and Sengupta (2013) provide a costly signaling model of premarital confinement in which women are supposed to have an underlying trait of "docility," and the marginal cost to parents of confining a docile child is lower than the cost of confining a non-docile child. Thus, confinement itself is taken to be a credible signal of docility, which husbands in turn take as desirable in itself or indicative of the cost of future fertility controls.

level of sanction for individual episodes of transgression and can thereby credibly distinguish themselves from groups that contain a higher proportion of transgressors.

A Model of Purification-Type Violence

Purification violence occurs in situations where there is the potential for mutually beneficial cooperation in equilibrium, but cooperation is asymmetrically risky. We model this with a game involving two players: the family of a prospective bride (player 1) and the family of a prospective husband (player 2). The prospective bride may or may not be "chaste" (compliant with the prevailing norms of sexual behavior), which is the high-quality type for the purpose of this game: the other player obtains a higher payoff from interacting with this type.[9] We treat this as given exogenously by nature, with probability x that she is chaste. Before the husband's family makes a decision about whether to enter into a marriage with her, there is some chance that an *unchaste* bride's status will be revealed to the husband's family, in which case they will certainly reject the marriage. A chaste family obtains payoff 1 from marriage, and both types obtain payoff 0 from rejection. The expected payoff to a low-quality family from obtaining a marriage is $0 < b < 1$, representing the risk that the marriage will fail because of a subsequent revelation. The husband's family will receive payoff 1 if they agree to a marriage with a chaste bride, a payoff of $0 < a < 1$ for rejecting the marriage, and payoff 0 for agreeing to a marriage with an unchaste bride.

In this setting, we focus on a subgame in which a transgression has been revealed about a cousin or sibling, shaming the entire family and jeopardizing the planned marriage of the prospective bride. The bride's family is motivated to send a signal, provided it will induce the husband's family to agree to a marriage. The signal is to punish the transgressor at cost $c > 0$. Player 1's strategy is therefore characterized by the probability they will send the signal if the prospective bride is high-quality (α_H) and the probability that they will send the signal if the prospective bride is low-quality (α_L). The husband's family's strategy is characterized by the probability that they will accept the proposed marriage having received the signal (β_S) and having not received it (β_N). The game is depicted in extensive form in Fig. 3. Equilibria are formally derived in the ESM.

One key parameter is the probability, for low-quality families, that their type will be revealed before the marriage. This determines the difference in expected benefit between low- and high-quality groups. If the probability of revelation is low, there will be a relatively small difference in expected benefit for the two types (b will be close to 1), and it will be correspondingly more difficult to find a level of signaling cost that sustains a separating equilibrium. More likely a pooling equilibrium will arise in which all senders adopt the same signaling strategy, regardless of type.[10] A second important factor is the relation between x and the payoff of unconditional acceptance by the receiver. We focus our analysis on cases where $x < a$: this is a relatively adverse environment for husbands' families because it means that, in the absence of any information, the optimal strategy for the groom's family is to reject marriage.

---

[9] In typical cases, the groom's family won't just be concerned with the chastity of the prospective bride, but of the entire extended family. For ease of presentation, we omit this complication below.

[10] This suggests a possible avenue for policy interventions: in some societies, such as Turkey, so-called virginity testing is widely reported (Parla 2001; Pelin 1999). Stronger privacy rights, or otherwise making such practices more difficult, may make honor violence less worthwhile as a signal.
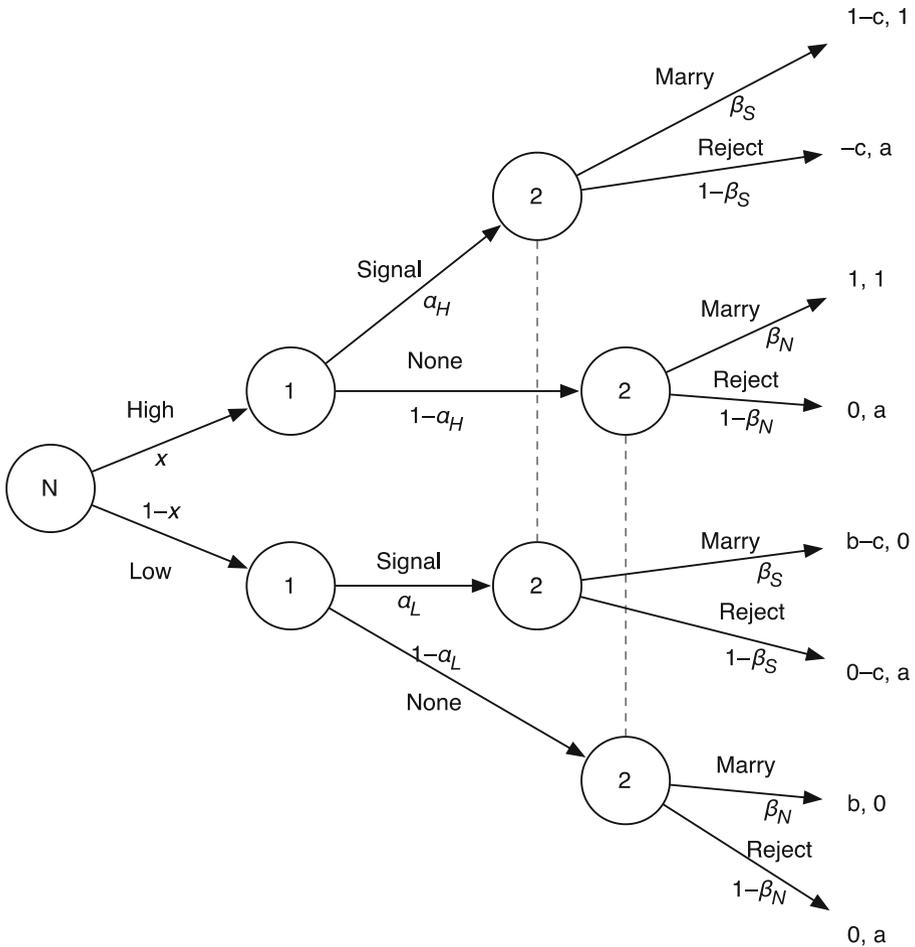
Fig. 3 The purification honor game. $0 < b < 1$; $0 < a < 1$; $c > 0$

In this model, honest signaling can stabilize because of the differential benefit that high- and low-quality types stand to obtain from the receiver. The key to achieving a separating equilibrium is that the signal must be sufficiently costly: it must be greater than b but less than 1 in order to price low-quality families out of sending the signal, while still making it worthwhile for high types. As in the case of violence associated with revenge-type norms, social norms within the family appear to buttress the strategic incentives to signal. Whereas all members of a family may benefit from sending a costly signal, it may be in no individual's interest to do so. A collective that has social norms that ensure the collective defense of honor will be much more likely to overcome this collective action problem. But by virtue of that strong commitment to honor, it will also be more likely that the family is of high quality. A daughter whose brothers may kill her for being seen unchaperoned with a male is much more likely to comply with the norms of chastity. So, there is a similar potential for a positive feedback loop for purification-type honor norms, as depicted in Fig. 2.

One further factor that might be relevant to understanding why the cost of signaling honor is often so high is the *competitive* nature of marriage and status in these honor societies. Good marriages are essential to secure important positional goods. One doesn't merely want to marry into a wealthy family, one wants to marry into a family *wealthier* than one's peers. Where there are positional goods at stake, there is reason to expect an arms race (Frank 2005). Indeed, Gideon Kressel et al. (1981) argues that honor killings are much more likely in socially mobile middle-class families than in families that are relatively stable at the top or bottom of the social hierarchy (see also Mackie 1996).[11]

This model explains (*i*) why it can be valuable to impose costs on members of one's own group (including one's family) and (*ii*) why those costs must be high to function as credible signals. In so doing, this model explains the puzzling features of violence associated with purification-type norms that seem to militate against the rationality or efficiency of the practice. As with violence associated with revenge-type norms, we expect to see these behaviors when the conditions of the model hold. Where there are large informational asymmetries between parties to an exchange, where quality is difficult to evaluate before the exchange, and where the adjudication for "fraudulent" exchange is corrupt, nonexistent, or costly, we should expect to see purification-type norms arise and persist.

Dueling as Purification Violence

The duel was a relatively ritualized form of honor violence, widespread in Europe during the sixteenth, seventeenth, and eighteenth centuries. Outwardly, duels appear to most closely resemble revenge-type honor violence because they are retaliatory conflicts occasioned by an insult or slight of some sort. This fits poorly with our prediction, however, that revenge-type norms are common in societies that face deterrence problems. The European nations where dueling was common had relatively robust property rights, especially for members of the aristocratic classes, who were the primary participants in dueling.

A model proposed by Douglas Allen and Clyde Reed (2006), however, suggests that despite its outward appearance, dueling much better fits the paradigm of purification. Dueling acts as a sort of screening mechanism to determine trustworthy individuals with whom to engage in productive exchange. Their central idea is that duels were a costly signaling device. Those with large amounts of social capital (high-quality individuals) stood to make sufficient gains through political patronage so it was rational to signal their quality by exposure to the risk of death or serious injury. Individuals with less social capital who wished to infiltrate the political class stood to gain less than those with more political capital and so would not be willing to pay a commensurate signaling cost and would be screened out. The latter individuals would decline to duel, thus leaving the political arena free for genuine aristocrats. Allen and Reed (2006) observe that their model explains the decline of dueling coinciding with the end of the era of political patronage and the emergence of a meritocratic bureaucracy. It also explains ritualized features of dueling: the role of seconds, the use of inaccurate weapons, and the extreme triviality of the insults that triggered many duels, all of

---

[11] It must be conceded, however, that it is not clear what form typical honor violence takes, and hence what exactly the distribution of behavior is that demands explanation. Other sorts of honor violence may serve a similar role but are reported less readily, such as physical mutilation and incapacitation (Niaz 2003). Also, families may attempt to conceal a child's transgression, rather than publicly punish it.

which suggest it was not an institution designed to accurately distinguish different degrees of physical danger that duelists posed (as hypothesized revenge violence would support), but to impose a relatively symmetric lottery on the participants.

Alternative Models of Honor Violence

Perhaps the model most similar to ours is developed by Richard McElreath (2003). He models the evolution of honor as the result of an evolutionary Hawk–Dove game in which strategists can decide whether to play hawk or dove in interactions based on the "reputation" of their partner, either "tough" or "weak," which is determined by their play in past encounters. In certain situations, the "tough" strategy (aggress against weak agents but defer to other tough agents) can become evolutionarily stable. We see McElreath's model as complementing ours by explaining, at a more basic level, how there can be selective pressure toward discrimination of reputational information and reputation-contingent strategies. His account explains why there might be preexisting but stable heterogeneity within the population with respect to predatory and retaliatory tendencies. Our model, using agents with greater cognitive sophistication, suggests how this heterogeneity may be reflected in more flexible signaling systems, and why we might then expect social norms to amplify a collective's commitment to honor norms and honor-based violence.

The McElreath model, however, does not address cooperative settings, nor does it explain in-group sanctioning. Hence it cannot explain purification honor violence. A better competitor, in this respect, is the "club goods" theory of religious and violent organizations (Chen 2010; Iannaccone 1992; Iannaccone and Berman 2006). This approach models extremist and violent groups as organizations that provide a kind of public good, a "club good," that is both rivalrous and excludable. In order to prevent free-riding from non-members, these organizations impose often severe costs on membership to show that one is a dedicated member of the group. Although some of the violence we discuss in this paper is likely to be explicable in these terms, we suggest that there is good reason to think that both approaches are necessary for a complete explanation. The club good model complements our own in that our account assumes that collectives already exist and are important units for problems of deterrence and assurance. The club good model explains this precondition of our own theory. The club good model, however, cannot explain why some of the costly rituals of groups are invariably demonstrated to out-groups. A signaling theory explains this readily. Finally, our model generates more specific predictions about the types of situations in which we will see one type of honor violence or another.

The Persistence of Honor Violence

Although we propose that honor norms serve a function in some social and economic contexts, one might hope that participants in honor cultures will abandon their norms once circumstances change. Where property rights are secure, where familial alliances and marriages are less important, we would hope to see that the demand for violence as a signal of quality will diminish and honor norms will lose their hold.

Certainly, there is some evidence of the converse: that individuals who are thrown into situations that create an elevated demand for assurance and deterrence will readily adopt honor norms. The study of prison gang organization by David Skarbek (2014) affords a compelling natural experiment in this regard. According to Skarbek, the enormous expansion of US prison populations in the late twentieth century rendered unworkable the earlier systems of norms that required small communities with frequent repeat encounters. Consequently, organized prison gangs have developed to provide protection to inmates and to provide warrant to outsiders that debts will be paid by their members. These functions are evidently addressing the deterrence and assurance problems, and many of the norms of prison gangs appear to be direct analogs to what we have called revenge and purification honor norms.

It is a mistake, however, to think that a norm will inevitably disappear once it ceases to become functional (Elster 1990). Two examples in particular illustrate that honor norms, once adopted, are relatively resistant to extinction, and thus regardless of whether they are ultimately social or moral/personal they can be internalized to the extent that individuals follow them in an unconditional way: (*a*) the evidence amassed for the vestigial persistence of a culture of honor in the US South, inherited initially from European migrants and then enduring through decades of significant economic and social change (Grosjean 2014; Nisbett and Cohen 1996), and (*b*) the observation that a number of honor killings have occurred in Western countries among migrants from South Asia or the Middle East (Chesler 2010). This may also be due to the increased fidelity and range of the honor signals in our globalized world, whereby family and clan members can communicate instantaneously through electronic means. There is also significant evidence that those raised in honor cultures are more likely to take gratuitous risks than others who have not internalized those norms and that this fact has a significant effect on their mortality (Barnes et al. 2012). Whether this is the result of a norm-psychology co-evolution or not is unclear.

Our model suggests another causal pathway by which a costly, inferior practice of honor might arise: via a path-dependent dynamic within the signaling game (Brennan et al. 2013:189–91). A signal may be optimal in one environment, where cooperation is risky and the introduction of costly signals enables more cooperation than would otherwise occur. But once those signals are widely understood and expected, if the environment becomes less risky, the signaling behavior may become an inefficient drag on cooperation. Given the positive feedback loop we noted earlier (Fig. 2) and the existence of social norms regulating honor, it is possible for these norms to remain stable even as the environment changes. We observe this in the purification-type honor model. When the expected payoff of unconditional acceptance is low, the separating equilibrium is a Pareto improvement. But if the risk becomes lower, the separating equilibrium is favorable to husbands but disadvantageous to brides. This nicely captures part of the situation with respect to honor killing. A costly signal was mutually beneficial in a high-risk environment of extreme polygyny, but in the lower-risk environment that now exists, it is gratuitous—not only is it an injustice against the individual victims, but it is also harmful to the interests of families who wish to contract marriages for their daughters.[12]

---

[12] Kim Sterelny (2007) argues that marriage practices are in general much less likely to be subject to adaptive cultural group selection because there is limited opportunity to experiment with different marital norms within a human lifespan, and marriage practices are not modular: they are embedded in larger networks of social practices. The larger complex may be subject to adaptive pressures, but not the component parts.

One possible avenue for disrupting these norms is by playing on the relationship between the individual and the group with relation to honor norms. Our models are designed to represent the incentives of groups (families, clans, and the like), but groups are made of individuals, and we predict significant conflict between the interests of individual and group. It may be highly contrary to individual interests to carry out an act of retaliation for the upholding of group honor, for instance. And even more obviously, the victim of an honor killing is hardly adequately compensated by the fact that her death may improve the marriage prospects of her cousins. Consequently, we have conjectured that groups employ social norms to enable them to harness the agency of individuals. These cleavages between group-level benefits and individual costs suggest an avenue for destabilizing honor norms. By reducing the group-level benefits (i.e., efficiency) of the honor norms and by highlighting or increasing the individual cost, honor norms will be less attractive solutions.

We have already identified that in some cases, honor cultures depend on a system of social conventions that might themselves be inefficient. The notion of "quality" that is relevant in purification signaling can be determined by social norms that have no intrinsic value and in fact require significant costs to maintain. The norms of chastity and modesty that deprive women of educational opportunities, limit their movement and employment, constrain their choice of dress, and so forth, are all burdens that women in particular, and indeed societies at large, would be better off without. Rather than being concerned primarily with the honor violence that is parasitic upon these notions of quality, it may be more helpful to try to undermine the quality conventions directly, as with the demise of footbinding.

References

Allen, D. W., & Reed, C. G. (2006). The duel of honor: Screening for unobservable social capital. *American Law and Economics Review, 8*(1), 81–115.

Appiah, K. A. (2011). *The honor code: How moral revolutions happen*. New York: W. W. Norton.

Barnes, C. D., Brown, R. P., & Tamborski, M. (2012). Living dangerously: Culture of honor, risk-taking, and the nonrandomness of "accidental" deaths. *Social Psychological and Personality Science, 3*(1), 100–107.

Benard, S. (2013). Reputation systems, aggression, and deterrence in social interaction. *Social Science Research, 42*(1), 230–245.

Benard, S. (2015). The value of vengefulness: Reputational incentives for initiating versus reciprocating aggression. *Rationality and Society, 27*(2), 129–160.

Bicchieri, C. (2002). Covenants without swords: Group identity, norms and communication in social dilemmas. *Rationality and Society, 14*(2), 192–228.

Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms*. Cambridge: Cambridge University Press.

Bicchieri, C. (2016). *Norms in the wild*. Cambridge: Cambridge University Press.

Black-Michaud, J. (1975). *Cohesive force: Feud in the Mediterranean and the Middle East*. Oxford: Basil Blackwell.

Boehm, C. (1986). *Blood revenge: The enactment and management of conflict in Montenegro and other tribal societies*. Philadelphia: University of Pennsylvania Press.

Brennan, G., & Pettit, P. (2006). *The economy of esteem: An essay on civil and political society.* Oxford: Oxford University Press.

Brennan, G., Eriksson, L., Goodin, R. E., & Southwood, N. (2013). *Explaining norms.* Oxford: Oxford University Press.

Chen, D. L. (2010). Club goods and group identity: Evidence from Islamic resurgence during the Indonesian financial crisis. *Journal of Political Economy, 118*(2), 300–354.

Chesler, P. (2010). Worldwide trends in honor killings. *Middle East Quarterly, 17*(2), 3–11.

Chesler, P., & Bloom, N. (2012). Hindu vs. Muslim honor killings. *Middle East Quarterly, 19*(3), 43.

Demetriou, D. (2014). What should realists say about honor cultures? *Ethical Theory and Moral Practice, 17*(5), 893–911.

Doris, J. M., & Plakias, A. (2008). How to argue about disagreement: Evaluative diversity and moral realism. In W. Sinnott-Armstrong (Ed.), *Moral psychology, Vol 2. The cognitive science of morality: Intuition and diversity* (pp. 303–331). Cambridge: MIT Press.

Efferson, C., Vogt, S., Elhadi, A., Ahmed, H. E. F., & Fehr, E. (2015). Female genital cutting is not a social coordination norm. *Science, 349*(6255), 1446–1447.

Elster, J. (1990). Norms of revenge. *Ethics, 100*(4), 862–885.

Fearon, J. D., & Laitin, D. D. (1996). Explaining interethnic cooperation. *American Political Science Review, 90*(4), 715–735.

Frank, R. (2005). Positional externalities cause large and preventable welfare losses. *The American Economic Review, 95*(2), 137–141.

Gambetta, D. (2009). *Codes of the underworld: How criminals communicate.* Princeton: Princeton University Press.

Getty, T. (1998). Handicap signalling: When fecundity and viability do not add up. *Animal Behaviour, 56*(1), 127–130.

Ginsburg, T. (2011). An economic interpretation of the Pashtunwali. *University of Chicago Legal Forum, 2011,* 89.

Gintis, H., Smith, E., & Bowles, S. (2001). Costly signaling and cooperation. *Journal of Theoretical Biology, 213*(1), 103–119.

Grafen, A. (1990). Biological signals as handicaps. *Journal of Theoretical Biology, 144*(4), 517–546.

Greif, A. (2006). *Institutions and the path to the modern economy: Lessons from medieval trade.* Cambridge: Cambridge University Press.

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

Hobbes, Thomas. 2012. Leviathan. Edited by Noel Malcolm. Clarendon Edition of the Works of Thomas Hobbes. Oxford: Oxford University Press, Leviathan.

Howard, J. A., & Gibson, M. A. (2017). Frequency-dependent female genital cutting behaviour confers evolutionary fitness benefits. *Nature Ecology & Evolution, 1.* https://doi.org/10.1038/s41559-016-0049.

Iannaccone, L. R. (1992). Sacrifice and stigma: Reducing free-riding in cults, communes, and other collectives. *Journal of Political Economy, 100*(2), 271–291.

Iannaccone, L. R., & Berman, E. (2006). Religious extremism: The good, the bad, and the deadly. *Public Choice, 128*(1–2), 109–129.

Johnstone, R. A. (1997). The evolution of animal signals. In J. R. Krebs & N. B. Davies (Eds.), *Behavioural Ecology: An Evolutionary Approach (fourth ed., pp. 155–78).* Wiley.

Kreps, D. M., & Wilson, R. (1982). Reputation and imperfect information. *Journal of Economic Theory, 27*(2), 253–279.

Kressel, G. M., Bausani, A., Ginat, J., Roger, J., Khazanov, A. M., Landau, S. F., Marx, E., & Shokeid, M. (1981). Sororicide/Filiacide: Homicide for family honour [and comments and reply]. *Current Anthropology, 22*(2), 141–158.

Kulczycki, A., & Windle, S. (2011). Honor killings in the Middle East and North Africa: A systematic review of the literature. *Violence Against Women, 17*(11), 1442–1464.

Kumar, V., & Campbell, R. (2016). Honor and moral revolution. *Ethical Theory and Moral Practice, 19*(1), 147–159.

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

Mackie, G. (1996). Ending footbinding and infibulation: A convention account. *American Sociological Review, 61*(6), 999–1017.

Mahdi, N. Q. (1986). Pukhtunwali: Ostracism and honor among the Pathan Hill tribes. *Ethology and Sociobiology, 7*(3), 295–304.

McElreath, R. (2003). Reputation and the evolution of conflict. *Journal of Theoretical Biology, 220*(3), 345–357.

Miller, W. (1990). *Bloodtaking and peacemaking: Feud, law, and society in Saga Iceland*. Chicago: University of Chicago Press.

Moritz, M. (2008). A critical examination of honor cultures and herding societies in Africa. *African Studies Review, 51*(2), 99–117.

Niaz, U. (2003). Violence against women in south Asian countries. *Archives of Women's Mental Health, 6*(3), 173–184.

Nisbett, R., & Cohen, D. (1996). *Culture of honor: The psychology of violence in the south*. Boulder: Westview Press.

Parla, A. (2001). The "honor" of the state: Virginity examinations in Turkey. *Feminist Studies, 27*(1), 65–88.

Pelin, S. Þ. (1999). The question of virginity testing in Turkey. *Bioethics, 13*(3–4), 256–261.

Rai, B., & Sengupta, K. (2013). Pre-marital confinement of women: A signaling and matching approach. *Journal of Development Economics, 105*, 48–63.

Ross, C. T., Strimling, P., Ericksen, K. P., Lindenfors, P., & Mulder, M. B. (2016). The origins and maintenance of female genital modification across Africa. *Human Nature, 27*(2), 173–200.

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

Selten, R. (1978). The chain store paradox. *Theory and Decision, 9*(2), 127–159.

Sessions, W. (2010). *Honor for us: A philosophical analysis, interpretation and defense*. London: Continuum.

Shafa, S., Harinck, F., Ellemers, N., & Beersma, B. (2014). Who are you calling rude? Honor-related differences in morality and competence evaluations after an insult. *Negotiation and Conflict Management Research, 7*(1), 38–56.

Skarbek, D. (2011). Governance and prison gangs. *American Political Science Review, 105*(4), 702–716.

Skarbek, D. (2014). *The social order of the underworld: How prison gangs govern the American penal system*. Oxford: Oxford University Press.

Skyrms, B. (2010). *Signals: Evolution, learning, and information*. Oxford: Oxford University Press.

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

Sterelny, K. (2007). SNAFUS: An evolutionary perspective. *Biological Theory, 2*(3), 317–328.

Thrasher, J., & Vallier, K. (2015). The fragility of consensus. *European Journal of Philosophy, 23*(4), 933–954.

United Nations Population Fund. (2000). Violence against women and girls: A human rights and health priority. *The State of World Population, 2000*, 25–30 https://www.unfpa.org/publications/state-world-population-2000.

Weiner, M. S. (2013). *The rule of the clan: What an ancient form of social organization reveals about the future of individual freedom*. New York: Farrar, Straus and Giroux.

John Thrasher is an assistant professor in philosophy and the Smith Institute for Political Economy and Philosophy at Chapman University. He works primarily on the relationship between individual rationality and social rules, especially in evolutionary and dynamic contexts.

Toby Handfield is associate professor of philosophy at Monash University. His current research is concentrated on topics at the intersection of evolutionary theory and philosophical ethics.