Runaway signals: Exaggerated displays of commitment may result from second-order signaling

Julien Lie-Panis*[a,b,c] and Jean-Louis Dessalles[b]

[a] *Institut Jean Nicod, Département d'études cognitives, Ecole normale supérieure, Université PSL, EHESS, CNRS, 75005 Paris, France*
[b] *LTCI, Télécom Paris, Institut Polytechnique de Paris, 91120 Palaiseau, France*
[c] *Université de Paris, EURIP Graduate School for Interdisciplinary Research, 75004 Paris, France*

December 27, 2022

Abstract

To demonstrate their commitment, members of a group will sometimes all engage in a ruinous display. Such widespread, high-cost signals are hard to reconcile with standard models of signaling. For signals to be stable, they must honestly inform their audience; for signals to be honest, their costs need only deter certain undesirable individuals. To explain this phenomenon, we design a simple game theory model, which we call the signal runaway game. In this game, senders can engage in *second-order signaling*. They can pay a cost to express outrage at a non-sender. In doing so, they draw attention to their own signal, and benefit from its increased visibility. Using our model and a simulation, we show that outrage can stabilize widespread signals and can lead signal costs to run away. Second-order signaling may explain why groups sometimes demand displays of commitment from all their members, and why these displays can entail extreme costs, as they frequently do during wartime.

Keywords: costly signaling; commitment displays; ritual; game theory

*Corresponding author; Email: `jliep@protonmail.com`; ORCID: 0000-0001-7273-7704

1 Widespread, high-cost displays

Membership in human groups often involves ritual behaviors which appear arbitrary and wasteful to the non-initiated, ranging from the embarrassment of hazing and the time-constraints of religious practice to the emotional and physical scarring of certain rites or recruitment devices (Atran & Henrich, 2010; Cimino, 2011; Densley, 2012; Sosis et al., 2007; Whitehouse & Lanman, 2014). These behaviors have been explained as displays of prosocial commitment (Bulbulia & Sosis, 2011; Gambetta, 2009; Irons, 2001; Sosis, 2003). In accordance with this explanation, individuals who expend more time and energy in ritual activities are on average more generous towards other group members (Ruffle & Sosis, 2006; Soler, 2012; Xygalatas et al., 2013), and are perceived as such (Power, 2017; Purzycki & Arakchaa, 2013).

Yet, ritual displays differ from the way signals are traditionally understood in a crucial manner; they involve most, if not all, of the members of a social group (Gelfand et al., 2020). Widespread costly displays run counter to theoretical expectations. When individuals all invest in the same signal (e.g. an initiation rite), the signal is dishonest (Gintis et al., 2001). If onlookers are unable to distinguish between participants, the ritual is uninformative; in theory, it should be abandoned. When individuals invest in different levels of signaling (e.g. in a lower-ordeal or higher-ordeal ritual, Xygalatas et al., 2013), the overall signal is honest, but net costly for the least committed (Dessalles, 2014). If individuals are unable to distinguish themselves from the bottom of the pack, they are better off opting out of the display entirely.

Our proposal is that *not* sending a signal can sometimes expose to more serious consequences than mere missed social opportunities. In certain contexts, non-senders will be exploited by senders, who may chastise them to make their own signal more visible. Widespread displays could then emerge out of a single motivation: advertising one's prosocial commitment, by any means necessary.

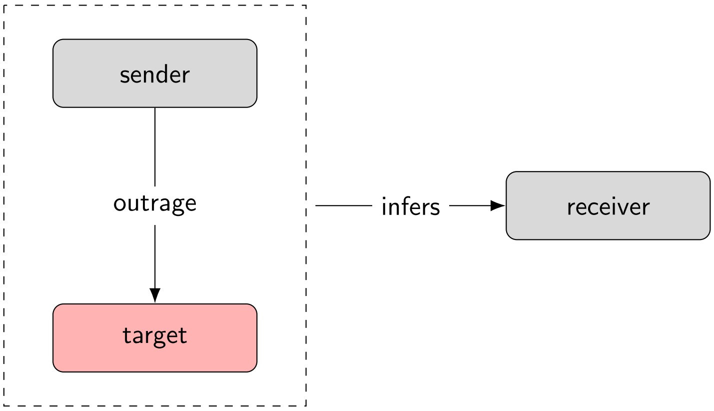


Figure 1: Outrage as a second-order signal. A sender can express outrage at a target who does not invest in the signal. When outrage is honest, receivers can infer that the sender has invested in the signal, even without having observed the sender's behavior directly. Outrage makes the sender's signal more visible. As a side-effect, the target is harmed.

More specifically, we argue that widespread costly displays can be propped up by moral outrage. Outrage can be a credible signal of moral behavior. To infer the moral quality of our partners, we sometimes use their propensity to express outrage (Jordan et al., 2017). Conversely, to advertise our investment in desirable behavior, we sometimes express outrage against those who unambiguously display undesirable behavior (Jordan & Rand, 2019); or even against those whose morality is merely ambiguous (Jordan & Kteily, 2022).

In the context of commitment displays, outrage can be thought of as a *second-order signal* — a signal about (the absence of) a signal (see Figure 1). When we publicly comment on others' perceived lack of investment in a display, we indirectly broadcast our own investment. In doing so, we increase others' incentive to display, and lay the groundwork for widespread signaling. To emphasize our own observance, we may for instance draw attention to those who secretly eat during a fast, and whose transgression may have otherwise gone unnoticed.

In this paper, we formally explore this hypothesis. We introduce a model, which we dub the 'signal runaway game', in which individuals may engage in first- and second-order signaling. Using our model and a computer simulation, we show that widespread costly displays may emerge endogenously, out of the motivation to advertise a socially desirable quality. We show that outrage can enable a step-by-step runaway process, leading individuals to gradually adopt costlier displays of commitment. Below, we outline the main elements of our model and simulation, and the main steps leading to our results (for a full characterization, see the Supplementary Information).

2 The signal runaway game

2.1 Baseline model

Commitment displays can be studied using the multi-player model introduced by Gintis, Smith and Bowles (2001), which we adapt. This type of model inevitably leads to a separating equilibrium in which only high-quality individuals pay the cost to send the signal.

We consider a large population where individuals are characterized by an unobservable quality q, which may take any value between 0 and 1, the minimum and maximum possible qualities. Individuals alternate between two roles, that of Signaler and Receiver. Signalers may pay cost $c_1(q)$ to send, depending on their quality q. Signaling is cheaper for high quality individuals: c_1 is a strictly decreasing continuous function of individual quality q which takes positive values. In the present context, individuals of higher quality can be thought of as individuals who are more committed to the group and/or its moral values, and whose commitment translates into an increased ability or willingness to invest in ritual signaling (e.g. because they expect to stay in the community for longer, and extract more social benefits from said community; Brusse, 2020).

Receivers choose a Signaler to follow. A signaling equilibrium occurs when they condition their choice on the signal; i.e. when Receivers pay to monitor others' signals, and follow a sender at random (rather than any individual). Receivers who monitor observe Signalers' behavior with probability $p_1 < 1$. Each time Signalers are chosen by a Receiver, they gain s.

Competition for followers leads to a separating equilibrium in which individuals send the signal when their quality is higher than a certain threshold $\hat{q}$, and do not send when it is lower. Let $\pi(\hat{q}) \equiv \mathbf{P}(q > \hat{q})$ be the fraction of individuals who send the signal. On average, Receivers observe a fraction $p_1 \times \pi(\hat{q})$ of senders, and choose one to follow. Signalers either do not send, and obtain nothing; or send, and are observed with probability p_1. On average, a Signaler recruits $\frac{p_1}{p_1 \pi(\hat{q}} = \frac{1}{\pi(\hat{q})}$ followers, earning s for each follower. $\hat{q}$ is the quality at which benefit and cost of signaling are equal, i.e. verifies:

$$c_1(\hat{q}) = \frac{s}{\pi(\hat{q})}. \tag{1}$$

For signaling to be stable, it must be honest. We obtain an evolutionarily stable strategy (ESS; Maynard Smith & Price, 1973) as long as Receivers benefit from following higher quality Signalers ($q > \hat{q}$) rather than lower quality signalers ($q \leq \hat{q}$), and that benefit exceeds the cost of monitoring. When monitoring is cheap, it is sufficient that the signal be prohibitively costly for individuals of minimum quality $q = 0$, i.e. that we have: $c_1(0) > \frac{s}{\pi(0)} = s$. In contrast, widespread signaling ($\hat{q} = 0$) is always uninformative, and can never be stable.

2.2 Outrage may sustain widespread costly signaling

The signal runaway game occurs when we introduce outrage into the previous baseline model. Signalers who send the signal may now pay c_2 to express outrage. Individuals who do not send cannot subsequently express outrage in our model; by assumption, outrage is a reliable indicator of signaling — a reliable second-order signal. We assume outrage increases the visibility of one's first-order displays. A sender who expresses outrage is observed with increased probability p_2 ($p_1 < p_2 < 1$).

Outrage is aimed in priority at non-senders in our model. When a Signaler expresses outrage, a target is selected at random among those individuals the Signaler observes opting out of the signal. That target is harmed, and loses h. A specific case occurs when the entire population sends the signal, and such targets are absent. In this case, we assume that outraged individuals may target ambiguous senders, i.e. individuals they do not observe sending the signal.

Signalers now compete to attract followers *and* evade others' outrage. Similarly to before, let us consider the case where Receivers condition on the signal, and Signalers send and express outrage when their quality exceeds a threshold $\hat{q} > 0$. As before, non-senders do not gain any followers, and miss out on average benefit $\frac{s}{\pi(\hat{q})}$. In addition, they risk becoming a target for the fraction $\pi(\hat{q})$ of outraged senders, with probability p_1. Outraged senders target one of the $p_1 \times (1 - \pi(\hat{q}))$ percent of individuals they observe opting out of the signal. Dividing, we deduce that non-senders lose on average: $\frac{\pi(\hat{q})}{1-\pi(\hat{q})} \times h$. $\hat{q}$ is the quality at which total benefit and cost of signaling are equal, and now verifies:

$$c_1(\hat{q}) + c_2 = \frac{s}{\pi(\hat{q})} + \frac{\pi(\hat{q})h}{1 - \pi(\hat{q})} \tag{2}$$

Outrage perturbs the typical signaling equilibrium, by increasing the incentive to signal. Sending the first- and second-order signal allows individuals to attract followers and evade others' outrage. When outrage is cheap ($c_2 = 0$), more individuals are pushed to send (the minimum bar $\hat{q}$ decreases).

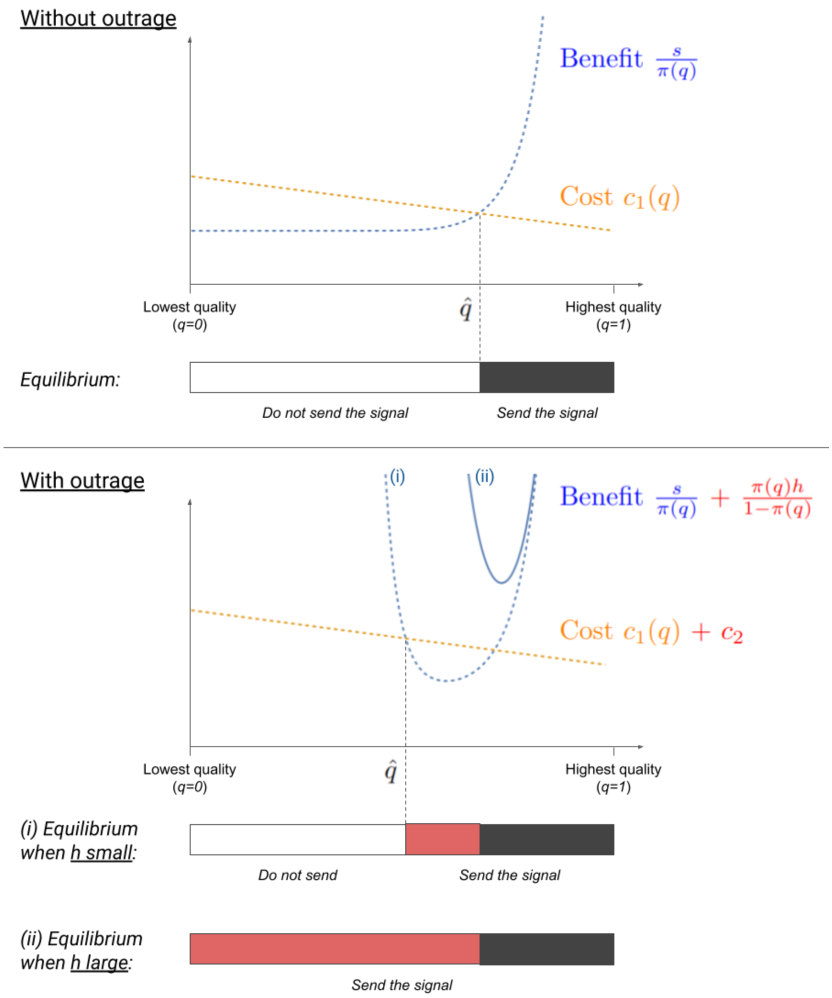


Figure 2: Effect of outrage on the signaling equilibrium. In the absence of outrage (top), a separating equilibrium is established at the threshold quality $\hat{q}$ which equalizes cost and benefit of signaling. Outrage increases the incentive to signal, as senders attract followers *and* evade others' outrage (bottom). (i) When harm h is low, we obtain another separating equilibrium, with a lower threshold quality; (ii) when harm is high, we obtain widespread signaling ($\hat{q} = 0$). For the purpose of illustration, we assume a linear cost function $c_1(q) = c_1(0)+q(c_1(1)-c_1(0))$, and that quality is normally distributed around $\bar{q} = 0.25$, with standard deviation 0.1. Other parameters: $c_1(0) = 2$, $c_1(1) = 1$, $s = 1$, $c_2 = 0.5$. In condition (i), we take $h = 0.01$; in condition (ii), we take $h = 0.1$ — with these parameter values, widespread signaling is obtained even with relatively small, but not minuscule, values of h.

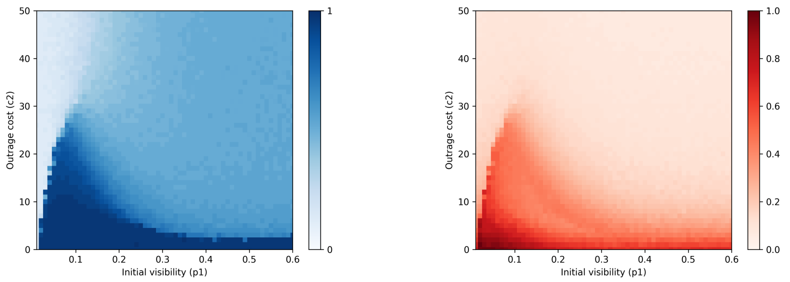
There are two possible outcomes, represented in Figure 2. First, when harm h is low, outrage introduces a small perturbation, and we retain a separating equilibrium. Second, when the consequences of being the subject of others' outrage are dire, outrage introduces a larger perturbation — and may push the population towards widespread signaling. We show that the minimum bar $\hat{q}$ decreases all the way towards 0 if:

$$c_1(0) + c_2 < s + 2\sqrt{hs} \tag{3}$$

Widespread signaling may then remain stable, even though it is dishonest. When $\hat{q} = 0$, the signal is uninformative for Receivers, and senders do not attract more followers than non-senders. Yet, any individual who attempts to save on the cost of signaling risks become the group's moral punching bag, by constituting a preferential, unambiguous target for others' outrage. We show that widespread signaling is stable when:

$$c_2 < \frac{(p_2 - p_1)h}{1 - p_2} \tag{4}$$

We implement our model into an agent-based simulation. Agents interact based on three flexible behavioral traits: their investment in a certain signal, their probability of expressing outrage at lesser senders, and of monitoring others' signals. Agents observe non-senders directly, with probability p_1, and indirectly via dyadic encounters with outraged partners. When initial visibility p_1 and the cost of outrage c_2 are small, agents learn to express outrage with high probability, and widespread signaling ensues (see Figure 3).



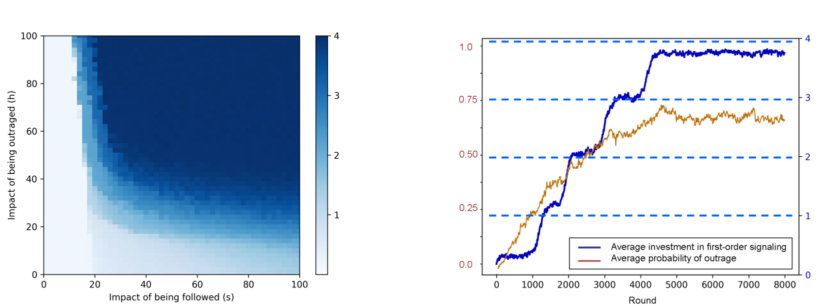
(a) Fraction of senders (b) Average probability of outrage

Figure 3: Simulation results, for one level of signaling. Agents' behavior in a given round is a function of three flexible traits: their investment in a certain signal, their probability of expressing outrage at lesser senders, and of monitoring others' signals. In the initial round, these traits are set at 0. With a small probability, agents may try out another value of the trait. The simulation and its parameter values are detailed in the Supplementary Information; code and figures are available from this website.
Left: fraction of senders after many rounds. Widespread signaling (dark blue) is obtained for low values of p_1 and c_2. Lighter blue colors represent mixed equilibria with a smaller fraction of senders. Right: average probability of outrage after many rounds.

2.3 Outrage may lead to exaggerated signal costs

When signaling is widespread, onlookers can no longer determine who are the top-quality individuals. To attract followers, these individuals may find it in their interest to create and adopt a new discrete signal level, requiring an additional investment of $\Delta c_1(q)$. Again, we assume Δc_1 is a decreasing function of individual quality q. Over-performers have every incentive to advertise their increased investment — e.g. by finding new targets of outrage. We assume they may now pay Δc_2 to express outrage at individuals who are observed sending at the lower level, and guarantee visibility $p_3 > p_2$; targets lose h. Similarly to before, individuals are pushed to increase their investment in the signal (they are prevented from decreasing their investment to 0 for the same reasons as before). We expect full escalation to the new signal level when:

$$\Delta c_1(0) + \Delta c_2 < s + 2\sqrt{hs} \tag{5}$$



(a) Average level of signaling

(b) Step-by-step runaway

Figure 4: Average investment in the signal after many rounds (left), and step-by-step runaway (right), for four evenly spaced levels of signaling. When harm h and benefit of being followed s are sufficiently high, agents learn to invest in the highest level of first-order signaling, and in high levels of second-order signaling (high probability of expressing outrage).

Outrage may lead a population to adopt exaggerated displays. We relaunch our simulation with several evenly spaced levels of signaling (proportional costs). Agents may now express outrage at non-senders and lower-level senders (whom they still observe directly and indirectly). When h and s are sufficiently large, outrage enables a step-by-step runaway process: individuals gradually learn to invest in the highest level of signaling (see Figure 4). This is in accordance with equation (5); when levels are evenly spaced, the marginal cost of signaling one level above is constant from one level to the next, and signal escalation may continue indefinitely. In reality, we expect marginal costs to increase at each step to infinity, as individuals are forced to miss out on increasingly important opportunities. The process will necessarily come to a halt. Eventually, high quality individuals will not benefit from creating a costlier display (and advertising it at the expense of others), and low quality individuals will prefer not to increase their investment, even if this means appearing relatively uncommitted.

3 Discussion

This paper offers a proof of concept for the existence of widespread costly displays. Our model is agnostic about any function the emerging behavior may serve at the level of the collective (e.g. encouraging group cohesion or cooperation; Atran & Henrich, 2010; Bulbulia & Sosis, 2011; Cimino, 2011; Durkheim, 2008; Gambetta, 2009; Irons, 2001; Whitehouse & Lanman, 2014; Xygalatas et al., 2013). Widespread signals are explained at the individual level. Outrage benefits senders, by making their signal easier to spot. We show that, under certain conditions, outrage is sufficient to generate widespread signaling, and escalating costs.

We consider signals which take discrete values. Our model applies for displays of commitment which categorize individuals (e.g. into participants of a high-ordeal ritual, of a low-ordeal ritual, and non-participants; Xygalatas et al., 2013), not when evaluations are based on a more continuous metric (e.g. time given to community work). This is a feature of the model, and not a bug. Though continuously-valued signals may emerge and remain stable (Grafen, 1990), outrage requires clear-cut comparisons. In some cases, committed individuals could design discrete displays precisely for that purpose.

We assume however that outrage is honest, in our model and simulation. Outrage is generally believed to be honest when hypocrites suffer sufficient retaliatory costs; yet, retaliation against hypocrites is subject to much variation (Sommers & Jordan, 2022). Further research should investigate the conditions under which outrage is more likely to be honest, and/or treated as such by onlookers; ensuring that it can function as a second-order signal.

Our model may help explain mandatory displays of commitment, such as rites of passage (see also: Cimino, 2011; Densley, 2012; Gambetta, 2009; Iannaccone, 1992). Outrage can create a positive feedback loop, and sustain uniform, and therefore uninformative, displays. The resulting behavior is a specific type of norm. In general, norms can emerge from a variety of positive feedback loops, such as those created by social punishment or benchmark effects (Young, 2015). In our case, uniform displays arise endogenously, from the motivation to advertise one's prosocial commitment to group members, via first- and second-order signaling (we do not need to assume non-senders are punished).

Our model may also help explain exaggerated displays of commitment, e.g. during wartime (see also: Sosis et al., 2007; Whitehouse, 2018). Times of crisis tend to favor expression of commitment over others (Hahl et al., 2018), and may provide the initial push enabling signal runaway. In extreme cases, the system is expected to stop at extreme levels of signaling and outrage, pushing individuals to ever greater lengths to avoid appearing uncommitted. A similar logic may be at play with witch hunts or other collective crazes which follow a self-fulfilling pattern (Lotto, 1994).

The present model is kept minimal. It needs to be completed to explain why many widespread signals remain stable without reaching extreme values, or why they may deescalate. Depending on the context, individuals may look for commitment to other groups or values. Signals and non-signals can change meaning (e.g. pacifism instead of cowardice, or closed-mindedness instead of dedication to the group).

Methods

Static analysis. To explore the conditions under which outrage may evolve, and lead to widespread signaling, we characterize all evolutionarily stable strategy (ESS) of the signal runaway game (for all details, see Supplementary Information).

Evolutionary simulations. To explore the conditions under which outrage may lead to widespread signaling and/or exaggerated signaling costs, and the evolution of strategies in a more realistic setting, we implement the model into an agent-based simulation (with one or several available signal levels). In the simulation, agents are characterized by a fixed quality, and three flexible features. They interact locally, based on their feature values at a given point in time. They learn optimal feature values by exploring the feature space, based on the outcome of these interactions.

The simulation is written in Python and based on the *Evolife* platform (for all details, see Supplementary Information). All programs are open source and available from the companion website, along with instructions for installation, figures, and chosen parameter values.

Acknowledgements

We thank A. Sijilmassi for feedback on a early version of the manuscript. This research was supported by funding from the EURIP Graduate School for Interdisciplinary Reseach, and from the Agence Nationale pour la Recherche (ANR-17-EURE-0017, ANR-10-IDEX-0001-02).

References

Atran, S., & Henrich, J. (2010). The Evolution of Religion: How Cognitive By-Products, Adaptive Learning Heuristics, Ritual Displays, and Group Competition Generate Deep Commitments to Prosocial Religions. *Biological Theory*, *5*(1), 18–30. https://doi.org/10.1162/BIOT_a_00018

Brusse, C. (2020). Signaling theories of religion: Models and explanation. *Religion, Brain & Behavior*, *10*(3), 272–291. https://doi.org/10.1080/2153599X.2019.1678514

Bulbulia, J., & Sosis, R. (2011). Signalling theory and the evolution of religious cooperation. *Religion*, *41*(3), 363–388. https://doi.org/10.1080/0048721X.2011.604508
_eprint: https://doi.org/10.1080/0048721X.2011.604508

Cimino, A. (2011). The Evolution of Hazing: Motivational Mechanisms and the Abuse of Newcomers. *Journal of Cognition and Culture*, *11*(3-4), 241–267. https://doi.org/10.1163/156853711X591242

Densley, J. A. (2012). Street Gang Recruitment: Signaling, Screening, and Selection. *Social Problems*, *59*(3), 301–321. https://doi.org/10.1525/sp.2012.59.3.301

Dessalles, J.-L. (2014). Optimal investment in social signals. *Evolution*, *68*(6), 1640–1650. https://doi.org/10.1111/evo.12378

Durkheim, E. (2008, June 15). *The Elementary Forms of Religious Life* (M. S. Cladis, Ed.; C. Cosman, Trans.; Abridged edition). Oxford University Press.

Gambetta, D. (2009). *Codes of the Underworld: How Criminals Communicate*. Princeton University Press.

Gelfand, M. J., Caluori, N., Jackson, J. C., & Taylor, M. K. (2020). The cultural evolutionary trade-off of ritualistic synchrony. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1805), 20190432. https://doi.org/10.1098/rstb.2019.0432

Gintis, H., Smith, E. A., & Bowles, S. (2001). Costly Signaling and Cooperation. *Journal of Theoretical Biology*, *213*(1), 103–119. https://doi.org/10.1006/jtbi.2001.2406

Grafen, A. (1990). Biological signals as handicaps. *Journal of Theoretical Biology*, *144*(4), 517–546. https://doi.org/10.1016/S0022-5193(05)80088-8

Hahl, O., Kim, M., & Zuckerman Sivan, E. W. (2018). The Authentic Appeal of the Lying Demagogue: Proclaiming the Deeper Truth about Political Illegitimacy. *American Sociological Review*, *83*(1), 1–33. https://doi.org/10.1177/0003122417749632

Iannaccone, L. R. (1992). Sacrifice and Stigma: Reducing Free-riding in Cults, Communes, and Other Collectives. *Journal of Political Economy*, *100*(2), 271–291. https://doi.org/10.1086/261818

Irons, W. (2001). Religion as a hard-to-fake sign of commitment. In R. M. Nesse (Ed.), *Evolution and the Capacity for Commitment*.

Jordan, J. J., & Kteily, N. S. (2022). People punish moral transgressions for reputational gain, even when they personally question whether punishment is merited. *Available at PsyArXiv*.

Jordan, J. J., & Rand, D. G. (2019). Signaling when no one is watching: A reputation heuristics account of outrage and punishment in one-shot anonymous interactions. *Journal of Personality and Social Psychology*, No Pagination Specified–No Pagination Specified. https://doi.org/10.1037/pspi0000186

Jordan, J. J., Sommers, R., Bloom, P., & Rand, D. G. (2017). Why Do We Hate Hypocrites? Evidence for a Theory of False Signaling. *Psychological Science*, *28*(3), 356–368. https://doi.org/10.1177/0956797616685771

Lotto, D. (1994). On Witches and Witch Hunts: Ritual and Satanic Cult Abuse. *The Journal of Psychohistory; New York*, *21*(4), 373–396.

Maynard Smith, J., & Price, G. R. (1973). The Logic of Animal Conflict. *Nature*, *246*(5427), 15–18. https://doi.org/10.1038/246015a0

Power, E. A. (2017). Discerning devotion: Testing the signaling theory of religion. *Evolution and Human Behavior*, *38*(1), 82–91. https://doi.org/10.1016/j.evolhumbehav.2016.07.003

Purzycki, B. G., & Arakchaa, T. (2013). Ritual Behavior and Trust in the Tyva Republic. *Current Anthropology*, *54*(3), 381–388. https://doi.org/10.1086/670526

Ruffle, B. J., & Sosis, R. (2006). Cooperation and the in-group-out-group bias: A field test on Israeli kibbutz members and city residents. *Journal of Economic Behavior & Organization*, *60*(2), 147–163. https://doi.org/10.1016/j.jebo.2004.07.007

Soler, M. (2012). Costly signaling, ritual and cooperation: Evidence from Candomblé, an Afro-Brazilian religion. *Evolution and Human Behavior*, *33*(4), 346–356. https://doi.org/10.1016/j.evolhumbehav.2011.11.004

Sommers, R., & Jordan, J. (2022). When does moral engagement risk triggering a hypocrisy penalty? https://doi.org/10.31234/osf.io/w23ec

Sosis, R. (2003). Why aren't we all hutterites?: Costly signaling theory and religious behavior. *Human Nature*, *14*(2), 91–127. https://doi.org/10.1007/s12110-003-1000-6

Sosis, R., Kress, H. C., & Boster, J. S. (2007). Scars for war: Evaluating alternative signaling explanations for cross-cultural variance in ritual costs. *Evolution and Human Behavior*, *28*(4), 234–247. https://doi.org/10.1016/j.evolhumbehav.2007.02.007

Whitehouse, H. (2018). Dying for the group: Towards a general theory of extreme self-sacrifice. *Behavioral and Brain Sciences*, *41*. https://doi.org/10.1017/S0140525X18000249

Whitehouse, H., & Lanman, J. A. (2014). The Ties That Bind Us: Ritual, Fusion, and Identification. *Current Anthropology*, *55*(6), 674–695. https://doi.org/10.1086/678698

Xygalatas, D., Mitkidis, P., Fischer, R., Reddish, P., Skewes, J., Geertz, A. W., Roepstorff, A., & Bulbulia, J. (2013). Extreme Rituals Promote Prosociality. *Psychological Science*, *24*(8), 1602–1605. https://doi.org/10.1177/0956797612472910

Young, H. P. (2015). The Evolution of Social Norms. *Annual Review of Economics*, *7*(1), 359–387. https://doi.org/10.1146/annurev-economics-080614-115322