Signaling theories of religion: models and explanation

Carl Brusse

To cite this article: Carl Brusse (2020) Signaling theories of religion: models and explanation, Religion, Brain & Behavior, 10:3, 272-291, DOI: 10.1080/2153599X.2019.1678514

To link to this article: https://doi.org/10.1080/2153599X.2019.1678514

Published online: 27 Dec 2019.

Submit your article to this journal

Article views: 94

View related articles

View Crossmark data

Citing articles: 1 View citing articles
Signaling theories of religion: models and explanation

Carl Brusse \(^{a,b}\)

\(^{a}\)Department of Philosophy and Charles Perkins Centre, University of Sydney, Sydney, Australia; \(^{b}\)School of Philosophy, The Australian National University, Canberra, Australia

ABSTRACT

The signaling theory of religion has many claimed virtues, but these are not necessarily all realizable at the same time. Modeling choices involve trade-offs, and the available options here have not traditionally been well understood. This paper offers an overview of signaling theory relevant to the signaling theory of religion, arguing for a narrow, “core” reading of it. I outline a broad taxonomy of the choices on offer for signaling models, and examples of how previous and potential approaches to modeling religious signaling meet or fail to meet the initial promise of the theory. A pluralist approach to religious signaling seems possible, but this would require a high level of detail and specificity with respect to both formal models and target systems.

ARTICLE HISTORY

Received 5 July 2018
Accepted 4 December 2018

KEYWORDS

Religion; signaling; evolution; modeling; explanation

Introduction

Religious signaling theory (RST) offers a joint, co-evolutionary explanation of two features of human society which are otherwise surprising when seen through the lens of evolutionary theory. One is our high level of social trust and cooperation, often in the absence of clear benefits. The other is the prevalence of public religions, whereby people submit to seemingly arbitrary demands and restrictions; again, without obvious immediate material reward. Both are anomalous. Even if human beings had been preconfigured for sociality and religion, the grim logic of game theory and selection at the individual level predicts that they should wither on the vine or be nipped in the bud. But these features appear to be widespread and stable in traditional human societies, and RST is one attempt to provide a more sophisticated naturalistic explanation for them (Bulbulia & Sosis, 2011; Cronk, 1994; Irons, 2001).

However, the position of RST in the literature is complicated. What constitutes a signal? Must signals be honest to proliferate? Must they be costly? Researchers with differing influences have answered or avoided these questions in different ways, meaning that labels like “signalling theory” or “costly signalling theory” are not always used to refer to the same sets of views and approaches. Imperfect congruence at the conceptual or terminological level means that despite promising approaches toward testing ostensive predictions of RST (Power, 2017; Sosis, 2003; Sosis & Bressler, 2003), it is not always clear which theories are being tested and how to tease them apart.

To take one example, Joseph Henrich’s theory of credibility-enhancing displays or CREDs (2009) is seen by many—including Henrich—as an alternative and rival to RST, while others have described it as an extension of it via cultural learning (Bulbulia & Sosis, 2011). CREDs are costly displays of sincere belief which increase the transmission of the relevant beliefs/practices, independently of how veridical or individually beneficial those beliefs might be. In a natural-language sense of the word, this is clearly a “signal” of some sort, but is it of the same kind as in other signaling theories?
The question here is not about the semantics of “signal”, but rather about the principled carving up of theoretical space. Is there only a disjunctive collection of “signalling” explanations (merging smoothly into other forms of explanation in the periphery), or is there a scientifically significant demarcation to be made? If so, how?

I argue that there is indeed a central notion to be pinned down which deserves its own label. The main contribution of this paper will be an abstract reappraisal of this notion from the inside out—focusing first on abstract signaling models and only then on their application to religion. This is intended as a contrast and compliment to more traditional surveys of the literature (e.g., Bourrat, 2015; Shaver & Bulbulia, 2016). The first goal is to sketch out a “core” understanding of RST; outlining distinctive commitments and explanatory virtues. This will (I hope) be ecumenical and recognizable enough to be an umbrella concept for most putative signaling theories of religion, while also serving as a reasonable basis for demarcation. In the middle and final sections I take a deeper look at the available options for RST: different models of signaling, and templates for their interpretation onto real-world religious signaling phenomena.

The signalling theory as an evolutionary explanation

The literature on signaling theory in religion has a variety of influences and is far from monolithic. Often going under the name of the “costly” or “honest” signaling theory of religion, the view is most closely associated with the work of Joseph Bulbulia, Richard Sosis, and co-authors (Bulbulia, 2004a; Bulbulia & Sosis, 2011; Shaver & Bulbulia, 2016; Sosis, 2003; Sosis & Alcorta, 2003), with earlier articulations by anthropologist William Irons (1996, 2001) and others in the context of the biological evolution of moral traits and institutions (Alexander, 1987; Cronk, 1994). One clear precedent is the theory of signaling in animal communication, including signals as manipulation (Dawkins & Krebs, 1978; Krebs & Dawkins, 1984), and the so-called handicap principle of costly signaling (Grafen, 1990; Zahavi, 1975; Zahavi & Zahavi, 1997), with signaling in biology (especially ecology) emerging as a complex field of study into its own right (Biernaskie, Perry, & Grafen, 2018; Hebets et al., 2016; Hebets & Papaj, 2005; Hurd & Enquist, 2005). A relevant parallel literature on costly signaling also exists in economics, beginning with Michael Spence’s Nobel prize-winning model of job-market signaling (Spence, 1973), and including economic analyses of organizational benefits and costly religious demands (e.g., Iannaccone, 1992). Another oft-cited influence is Robert Frank’s theory of emotional displays as evolved signals, which strategically advertise commitment to specific courses of action (Frank, 1988). As these influences suggest, the development of RST has not been linear, with multiple authors from different disciplinary and methodological backgrounds arriving at it independently (Sosis, 2005).

The theories arrived at are also quite diverse with respect to how signaling supposedly works, and over what scope and timescale. For example, Sosis sees religious signaling as pro-social mechanism for large-groups (with reputational mechanisms dominating in small groups), while for Bulbulia it acts on a local, observable scale (Bulbulia, 2010; Sosis, 2005). And while the biological and formal signaling models which act as a precedent are also typically based on population thinking (i.e., modeling signaling via dyadic interactions within populations of more-or-less interchangeable individuals), more complex verbal models with social/network configurations have also been proposed e.g., top-down, one-to-many religious leadership or “charismatic” signaling (Bulbulia, 2010; Shaver, Fraser, & Bulbulia, 2016), and ritualized, aposematic signaling of power and brutality (either top-down or between-groups) (Bulbulia, Fraser, Watts, Shaver, & Gray, 2017). Such different approaches, along with differing appeals to cognitive processes and other causal mechanisms, place differing emphasis on religious mental content on one hand and social-explanatory targets on the other, e.g., the evolution of basic co-operation, social order, or power hierarchies and inequality. There is, in other words, a rich and heterogeneous literature that I will not fully survey.
One dimension of difference that I will bracket off until the second section of the paper is how signals are kept honest. Many discussions about signaling theory in religion begin with examples of costly or intense religious practices, and appeal to signaling models originating elsewhere (in biology, or sometimes economics) which seem to fit them, such as costly signaling, and/or index signals. I will be arguing that the modeling picture is much more complicated than this, and in any case secondary to the business of characterizing RST itself.

In this first section, I focus instead on a simple (co-)evolutionary schema as the characteristic core of the RST, with more sophisticated theories in the literature seen as variations on that theme. This will also help to isolate why we might be interested in a signaling theory of religion in the first place: as explanatory package deal in which human religious and cooperative traits co-evolved, analogous to co-evolved trait-pairings in biology (such as the striking morphology of flowering plants on one hand and the foraging behavior of pollinating insects on the other). This schema has several features which make it particularly interesting in the abstract, which I will develop along the way.

**Mechanism, endogeny, and scalability**

The RST schema can be illustrated by way of comparison with the Big Gods theory (Norenzayan, 2013; Norenzayan et al., 2016), though the contrast should not be taken as adversarial. The Big Gods theory has been developed via several lines of evidence and a sophisticated combination of theoretical mechanisms (including CREDs and cultural group selection), and which also articulates a co-evolutionary relationship between religion and cooperation. For current purposes, I will focus on its distinctive claim: that beliefs in supernatural sources of “cosmic justice” serve to upregulate pro-social behavior among believers, despite costs that might be associated with those beliefs.

In contrast, the RST schema operates at a more functional and descriptively abstract level (in that first order psychological states and religious content are not central to its mechanism of action). Religious participation became a commitment signal for cooperation-inclined individuals to collectively condition their behavior upon. This allowed the positive assortment of cooperation, and the more reliable identification and exclusion of exploitative free-riders and bullies. Puzzling costs and demands become explicable as adaptive traits to better facilitate these commitment signals and/or ensure their reliability (see next section).

Both big gods beliefs and the RST mechanism can be seen as ways of improving on reciprocity as a way of solving the cooperation problem (Trivers, 1971), by “out-sourcing” the maintenance of cooperative behavior. The standard (prisoner’s dilemma) model of the cooperation problem is that each agent would do well if both cooperated with each other and poorly if both defected, but each would do best if they defected on a cooperating partner and worst if they cooperated and the other defected. For rational agents (or adaptively evolved behavioral strategies) mutual defection is the only expected outcome. Famously though, in iterated prisoner’s dilemma interactions (where agents may conditionalize behavior on past behavior of their partner), a “tit for tat” strategy which rewards cooperation with cooperation can be successful (Axelrod, 1984). Reciprocal altruism is a similar idea: agents conditionalize their behavior on the past behavior of their fellow group members and cooperate with those with an evidenced proclivity to cooperate. This facilitates the positive assortment of habitual co-operators and rewards cooperation.

But in practice, reciprocal altruism suffers from scaling limitations. The information required to assess every potential partner increases dramatically with increasing numbers of possible interaction partners, as does the monitoring required to gather that information. Some argue that these demands drove the striking evolution of human brain development and language (i.e., gossip) capabilities (Dunbar, 1998, 2009). In any case the theoretical limitation is clear: given some background rate of trustworthiness, the work that reciprocal altruism requires to stabilize social groups becomes problematic beyond a certain group size (for any cognitively limited creatures).
Big gods beliefs are one way to mitigate this, by improving the background rate. Believing that defection might provoke supernatural punishment should make it less likely as a behavior and thereby allow bigger societies to evolve. RST works instead by introducing religious signals as an alternative informational proxy on which to conditionalize cooperation, potentially more efficient than collated interaction history. Signals can be a good proxy in this way to the degree that they are (i) reliably correlated with likelihood of cooperation and (ii) highly observable. Ideally speaking, if reliable signals were directly observable “in the moment”, then the cap on group size might be eliminated entirely.

RST’s prospects will depend on how closely this ideal can be approached. For example, Bulbulia argues in favor of signal honesty, but sees the observability of religious commitment signals as limited and therefore favors a small-scale role for religious signaling (Bulbulia, 2010). Bulbulia has in mind discrete ritual displays, and there are indeed limitations here. Public ritual displays might be more observable than cooperation history, but the informational demands of observing and tracking prior ritual participation would be similar to those for reciprocal altruism. But there is also scope to consider persistent commitment signals, such as ongoing conformity to restrictive codes of dress, language, behavior, or association. These would be observable and understandable by cultural conspecifics even in the absence of any personal familiarity. Secondary signals (e.g., scarification or other earned status markers) might also reliably and persistently advertise prior commitment demonstrations. This is especially plausible if a religious community’s signifiers are well designed and well policed, so that faking a signifier of having “paid one’s dues” is either prohibitively difficult (e.g., artistically complex tattoos) or prohibitively risky (e.g., punishable impersonation of a high-status figure). Similar ideas will be returned to in section three, but for the moment (despite Bulbulia’s concerns) I will just state the following broad ambition for RST: that religious signaling can stabilize cooperation on a larger scale than reciprocal altruism, even in large social worlds.

**Features, function, & flexibility**

Several attractive features fall out of this abstract schematic recipe. It would operate at the same level of explanation as the problem it would address: individual-level fitness. It offers to by-pass the need for specific, personalized knowledge about potential cooperation partners; perhaps permitting arbitrary large societies to evolve. It also has relatively few moving parts in relation to its explanatory payoff: it doesn’t explicitly require any exogenous, independently arising cognitive biases, belief structures, or maladaptive traits (though these may be included in more detailed variations of it).

With regard to theory demarcation, the more endogenous an example of RST is (i.e., the less it causally depends on other specific features of ancestral humans or their environments), the closer that theory is to the core conception. On this characterization, CREDs theory would likely fall outside of RST proper, as religious CREDs are signals of religious sincerity rather than underlying value as a social partner, so do not facilitate fitness improvements. On the other hand, the “charismatic signalling” of Bulbulia and co-authors might be seen as a variation on the RST theme. It is in this sense that the core RST schema offers a way to help categorize and clarify the literature in a principled manner (though this will not be pursued further here).

It is also very much a functionalist approach. Backgrounding definitional issues (Harrison, 2006; White, 2018), the general explanatory target of the schema is the social phenomenon of religion: why beliefs and practices exist as socialized, locally-common features at the community level in virtue of their function. It therefore sits most naturally within a Durkeimian tradition focused on religious practices as central to group cohesion (Durkheim, 1912; Watson-Jones & Legare, 2016; Whitehouse & Lanman, 2014). RST can remain largely neutral on the content of the beliefs and practices in question because it is the signaling role of religion which makes it stable and ubiquitous, not any specific religious form or content (though this is often re-inserted by RST advocates e.g., via posited cognitive biases). The Big Gods theory again provides a useful contrast here, as it postulates that successful
religions are made so by spreading belief in powerful moralizing gods (or other sources of cosmic justice). In comparison, RST can avoid making predictions about any specific religious content and its correlation with success, with the putative functional role of religious signaling being multiply realizable.

This implies a high degree of diversity and flexibility. Like languages, signaling-driven religions could take on an arbitrary variety of forms—potentially explaining why the cross-cultural diversity of religious practices challenges definition & categorization efforts. It also allows potential applicability (via more specific RST mechanisms) to a variety of explanatory targets. These might include religious stability in different sized societies, and over different timescales: deep-time cognitive evolution, rapid cultural evolution, social transitions in the Pleistocene and Holocene, or specific application to historical eras (such as the axial age) or contemporary religious settings. Nothing prohibits (a) distinct cultural/ritual features from serving parallel/simultaneous signaling functions, or (b) single features having distinct signaling functions in concurrent effect.

In summary, this abstract, core notion of the RST is of a relatively parsimonious co-evolutionary explanation of religion and cooperation that could be highly scalable and flexible, depending on how the details are specified (who the agents are, models of signaling, realizer mechanisms etc). In principle, different possible versions of RST (including those in the literature) correspond to different ways of filling in those details, and of combining this core mechanism with others (such as exogenous cognitive influences and/or additional “training” effects from taking part in signaling rituals).

**Signaling theory for religious signaling**

Arguably the most important detail to fill in (for going from RST schema to testable theory) is how the honesty condition is satisfied: how religious signals come to be reliably correlated with value as a social partner.

Thus far I have avoided the usual tropes at this level of detail such as “costly signalling”, since there are several distinct honest signaling models to consider crucially, different models of honest signaling will meet the relevant requirements and desiderata to different degrees—and not all of them simultaneously. More comprehensive surveys of modeling possibilities are available (see for example (Zollman, 2013)), but it will be useful to consider the basics of signaling theory, and a basic taxonomy of the models available in the context of religious signaling.

**Signals and signaling systems**

Mirroring the heterogeneity of the literature on religious signaling, signaling theory itself was developed in such varied contexts as mate selection (Grose, 2011; Zahavi, 1975), ecology (Hebets et al., 2016), economics (Riley, 2001; Spence, 1973, 2002), and in formal modeling and philosophy of science (Grafen, 1990; Huttegger & Zollman, 2013; Skyrms, 2010). For sake of precisifying relevant terms and concepts though, I will largely follow (Maynard-Smith & Harper, 2003).

The first distinction to be made is between cues and signals. Cues are observable features of an organism which correlate with other features but did not originate for that reason. The criterion for a signal is communicative agency or teleology on the part of the sender as well—cues are merely observed, signals are also sent. Sender and receiver strategies which constitute signaling relationships can be passive or behavioral, and the result of natural selection or other mechanisms. For example, a species of poisonous frog might evolve distinctive coloring and/or a salience-enhancing display (a passive signal vs a distinct behavioral signal) because of a pre-existing coloring trait (a cue) that its predators learned to avoid. The co-evolutionary relationship between sender and receiver is important: once the predator began conditionalizing avoidance on the proto-signal, optimizing that coordination was in the interest of both parties.
This illustrates some key points. One is that real-world cases can admit of degrees. Cues can gradually evolve into signals and borderline and hybrid cue-signal cases are possible (during evolutionary transitions or otherwise). More importantly, joint evolutionary trajectories are important for explanation: a one-sided, cue-observer relationship does not explain where the cue came from. Explanatory signaling therefore requires convergent teleological backstories for both senders and receivers, via natural selection or some other mechanism.

Cues are where the sender doesn’t profitably participate in signaling relationship. The inverse of this is also possible. Consider the Ophrys speculum orchid, which attracts male Campitoscola ciliata wasps by mimicking the chemical signals of a sexually receptive female wasp, so much so that the males preferentially copulate with the flower rather than the actual females (Ayasse, Schiestl, Paulus, Ibarra, & Francke, 2003). Following Owren, Rendall, and Ryan (2010) and earlier precedents (Dawkins & Krebs, 1978; Krebs & Dawkins, 1984), cases like this are sometimes described as signals by influence or manipulation. But dishonest “signals” like this co-opt pre-existing responses, so (like cues) will not be stable unless those responses are stabilized by external evolutionary constraints or pressures.

This ambiguity is reproduced in the religious signaling literature. Both Cronk (1994) and Sosis and Alcorta (2003) discuss the likelihood and mechanics of manipulation via religious signaling. We should not quibble about how the terms should be used; indeed, the implication that signals and communication are honest by definition would sit poorly with natural language. What is important is that manipulative signals, cues, and co-adaptive signaling are not of the same explanatory kind. To avoid confusion, and following philosopher David Lewis (1969; Skyrms, 2010), we can identify a co-adaptive relationship like the frog-predator case as a game-theoretic signaling system. Focusing on signaling systems for the payoff-driven explanatory signaling we are interested in allows us to sidestep the semantic debate (though this also should not be confused with the “signalling system” terminology of e.g., Hebets et al., 2016).

The schema can therefore be updated: RST involves positing signaling systems, where religious signaling practices came about (or were stabilized) because of co-adaptation between participants and observers. This is payoff-driven, in that honest signals of social quality (e.g., social norm adherence) and sensitivity to those signals were both rewarded (e.g., by positive assortment with normatively compatible conspecifics). Once established, signaling systems become vulnerable to manipulative free-riders (all ecosystems have their parasites), which must be countered by some sort of honesty-biasing mechanism.

**Honesty in fakable and unfakable signals**

There are two canonical honesty-biasing mechanisms. I am calling these fakable signals and unfakable signals; though the nomenclature is not uniform. For example, unfakable signals are sometimes called hard to fake (because fakability admits of degrees), index (though “index” is sometimes used for any correlation between signal and underlying trait), constrained (Searcy & Nowicki, 2005) or performance signals (Hurd & Enquist, 2005). Even then (as we shall see also later) the use of labels like “hard to fake” and “index” can diverge (Cronk, 2005).

Unfakable signals reliably convey information about an underlying state or trait using a method that is unavailable (rather than just dis-preferred) in the absence of that underlying state. Maynard-Smith & Harper’s example is the roar of a stag in rut. The pitch, volume, and timbre of a roar are direct, reliable consequences of the stag’s size and strength. It is easy to intuit how this signaling system is incentivised: weak stags can use the information in the roar to avoid picking fights with strong stags, who can then save their strength for seeing off genuine rivals. If the connection between trait and signifier is not generative or otherwise diaphanous like this (or where alternative generation methods are available), the signals are more open to being sent by agents who lack the underlying qualities of interest, i.e., they are fakable.
Fakable signals and the problem of “costly signalling”

Without going too far into the details, Lewis’s original work was on systems of costless fakable signals as the solution to coordination games (and an idealization of language). In a coordination game the defining payoff structure is such that sender and receiver should prefer identical outcomes. Arbitrary conventions should emerge based on the separating equilibria of the game: stable pairings of conditional strategies that neither party have incentive to defect from (Brusse & Bruner, 2017). But these “cheap talk” equilibria disappear as interests start to diverge. If there are desirable “high-type” and undesirable “low-type” senders (from the perspective of receivers) who nevertheless both benefit from being treated as high types, then the low-types will have equal incentive to use any fakable “I’m a high type” signal, rendering it useless in the face of the cooperation problem.

Enter the misleading elephant(s) in the room: costly signaling and the handicap principle. The basis of costly signaling or handicap signaling systems (Grafen, 1990; Zahavi, 1975; Zahavi & Zahavi, 1997) is that the incentive to lie can be mitigated and reversed if the costs of sending signals are such that lying is no longer worth it. This has been appealed for explaining the cost and demandingness of religion (Alcorta, 2017; Bulbulia, 2004b, 2004a; Sosis, 2003; Sosis & Alcorta, 2003).

The handicap principle has a convoluted history which I won’t go into (see Grose, 2011) for an overview), but there are complexities to acknowledge. Most famously, it is not enough for signals to be costly: costs and benefits must be differentially distributed so that lies (but not honesty) are priced out of the evolutionary market. The general condition is that the average cost ($c_l$) of an “I’m a high type” signal by a low-type must be greater than the average benefit ($b_l$) of them being treated as a high type, while the average cost of the honest signal ($c_h$, for a genuine high-type) must be less than the relevant benefit ($b_h$), i.e.:

$$c_h < b_h, c_l > b_l$$

(1)

Assuming benefits are the same (i.e., $b_l = b_h = b$), this joint inequality condition can be simplified to a single inequality:

$$c_h < b < c_l$$

(2)

For receivers (and high-type senders), it is therefore strategic to coordinate on a signaling mechanism which imposes appropriately differential costs on the sender.

This differential cost requirement is just the first reason why simpler terms like “costly signalling” and “costly signalling theory of religion” are misleading. Another reason is that the relationship between costly signaling and fakable signaling is not one-to-one: unfakable/indexical signals can have unavoidable costs of production, usually dubbed “efficacy” costs in contrast to the “strategic” costs of (1) and (2) (Cronk, 2005; Lachmann, Számadó, & Bergstrom, 2001). And if signaling is stable (i.e., at equilibrium, where low types have learned that lying is not worth it) there may be no observable costs at all, because it is possible to meet condition (2) when $c_h$ is zero (Számadó, 2011). Finally, what looks like a costly signal might not be part of a signaling system at all: the costs demanded might instead constitute a practice of maladaptive ritual subjugation, or an extortive transfer of resources (e.g., from the younger generation to the old), or perhaps be better explained via another mechanism such as CREDS. The presence of costly, signal-like behaviors is therefore neither necessary nor sufficient for the handicap principle to be in play.

More generally, cost-talk can be problematically vague. Metabolic expenditures in simple microbial populations are ceteris paribus fitness costs (Birch, 2017), but energetic expenditures, painful experiences, or other intuitively “costly” signals might not translate into significant fitness costs at all. For example, a stag’s roar might leave it briefly overexerted, but whether that actually translates into a fitness (efficacy) cost depends on how soon it is challenged by another stag. The intuitive notion of a cost is broader than the narrower sense we need for a signaling explanation. Cost-talk
must therefore be sensitive to what Grose (2011) calls the currency of the cost: posited costs exist alongside benefits within the payoff structure of a signaling game, and must therefore be commensurable, additive, and significant in relation to each other. This is especially relevant for some interpretations of costs and benefits in religious signaling. For example, it is common to interpret the pain suffered by participants of extreme rituals as a cost of some kind (Power, 2017; Xygalatas et al., 2013). But if pain is a signal cost, then how is it traded off (as it must be, by hypothesis) against the putative economic/community-derived benefits? This is not unanswerable, but the question should not be ignored.

In a more direct challenge, the differential costs condition has been used to critique RST when understood as a handicap principle hypothesis. In laying out the CREDS alternative, Henrich argues that: “it is not clear why (in a fitness sense) it is more costly for nonbelievers to perform the costly requirements than for believers (more committed people)” (Henrich, 2009; a worry previously acknowledged by Sosis (2003). Indeed, the otherwise profitable time lost to a religious service is not sensitive to sincerity. And if tithes, arduous rituals, and other demands against one’s good standing in the religious community are tests of commitment, then they cannot then be pre-set higher or lower for the committed and the uncommitted. Of course, perceptions of signal cost might vary according to commitment level and this might have would-be free-riders balking at the price of entry. But relying on subjective costs breaks from the idea of RST as a fully co-adaptive theory—only objective costs drive adaptation. I will return to this proposal in the final section, as it is the approach favored in Sosis (2003). For the meantime, we can dub this the “differential objectivity” challenge for costly signaling.

But we have already seen that there are two reasons why the differential objectivity challenge might be less than fatal. The first is that we do not need to understand RST as a costly signaling or handicap principle hypothesis: signaling theorists could instead emphasize cognitively engaging, hard-to-fake signaling; and several have done so (Bulbulia, 2013; Bulbulia & Sosis, 2011; Shaver & Bulbulia, 2016). The second is that condition (2) assumes $b_l = b_h = b$, and if we relax that assumption we can accommodate equal signal costs $c_l = c_h = c$, providing a different condition is satisfied:

$$b_h > c > b_l$$

(3)

This “differential benefit” signaling option is well known, generally described either as complimentary to handicap signaling or subsumed under the same term (Bulbulia & Sosis, 2011; Johnstone, 1997; Maynard Smith, 1991; Murray & Moore, 2009). The evolutionary dynamics mirror that of differential cost signaling (Zollman, Bergstrom, & Huttegger, 2013), though it is arguably less well explored in terms of application to real-world cases.

In the final section I will expand on these ways of evading the differential objectivity challenge, with more exotic differential-cost models potentially available as well. But these modeling strategies need to be carefully disentangled from the more familiar notions of handicap and costly signaling.

Towards a taxonomy of religious signaling

We therefore have two more-or-less distinct classes of RST-relevant signaling system: (a) differential cost–benefit and (b) unfakable. Proper application of these models to real-world mechanisms should be sensitive to evolutionary timescale and ecological context, and empirically distinguishing them may be difficult without significant epistemic investment. Formally speaking, while they are not simply paraphrases of each other, fakable and unfakable signaling can also admit of degrees (Huttegger, Bruner, & Zollman, 2015). It should also be noted that the distinction itself has been challenged at an abstract level (Grafen, 1990) on the basis that all signals are fakable in principle over long enough an evolutionary timescale.

But for our purposes the distinction is “good enough”. And there are important (and potentially tractable) differences to investigate between different signaling regimes, especially when seen
through the lens of evolutionary trajectories. Hard-to-fake signals can have efficacy costs imposed by mechanistic necessity, but their costliness has no adaptive utility and evolution will do what it can to minimize them. In a costly-fakable signaling system, differential costs can have a positive function for evolution to upregulate. CREDs mechanisms in comparison predict costly display behavior propagating (via cognitive biases) but see no incentive for differential costs and benefits. So, while these explanations for costly displays might be difficult to tease apart (especially without historical data), there are real differences between them in terms of evolutionary dynamics and empirical prediction.

Returning to the explanatory desiderata from earlier in the paper, the two signaling system models also predict qualitative differences in signal characteristics. In principle, fakable signals can exhibit greater variation, versatility, and (perhaps) evolvability. In terms of cross-cultural diversity, they can be more “language-like”; with signals as more-or-less arbitrary symbols, floating freely of what they signify. Unfakable systems in contrast would be more restricted and “athletics-like”: ways of directly revealing/demonstrating underlying traits via constrained (presumably biological rather than cultural) connections between signal and trait.

**Figure 1** is a depiction of the space of signaling-like explanations. The costliness of signal-like behaviors cross-cuts with the fakable-unfakable distinction, with standard handicap signaling (differential cost) being a special case within a broader class of what is here labeled “differential cost–benefit” signaling. Shaded items are examples of signaling systems under partial conflict of interest (i.e., high and low sender types). Unshaded are Lewis-style costless fakable signals (cheap talk) with aligned interests, and flat cost signal mechanisms e.g., CREDs. These can convey information under specific conditions, but do not allow signaling systems in the adaptive, objective-payoff sense under conflict of interest. Also included under differential cost–benefit signaling are differential cost, differential benefit, and vulnerability cost models (which are variations on

---

**Figure 1.** Venn diagram classifying signaling models by costs/benefits and fakeability.
differential cost but with significantly different equilibria). Each of these should also be recognized as merely special cases of the general differential cost–benefit template (i.e., with either costs or benefits held fixed), and there is no reason why real, evolved signaling systems should approximate these instead of more general, hybrid forms satisfying condition (1).

One caveat to make in passing that the game-theoretic picture is more complex than the separating equilibria description given earlier. There are also partial pooling and hybrid equilibria, where communicative strategies coexist with or are mixed with unconditional strategies (Bruner, Brusse, & Kalkman, 2017; Kane & Zollman, 2015; Zollman et al., 2013). Unfakable signaling models have also been formally explored, with the introduction of efficacy costs sometimes producing partial-pooling equilibria (Jovanovic, 1982) or quashing the evolution of signaling altogether (Bruner, 2015). Such formal results often demonstrate the inadequacy of even educated intuitions regarding what signaling theory actually predicts.

The multiple modeling options and their complexities all mean that serious attempts to connect real-world, signal-like behaviors to signaling models will be demanding, especially in complex human societies where there are many traits of potential interest. For example, a display of wealth in the form of a religious donation might be a costly signal of social/religious commitment, but it might just as well directly signal for wealth or social dominance—it is after all valuable to know who the local elites are, and for the elites make themselves known as such, but elite status is famously distinct from prosocial commitment (Piff, Stancato, Côté, Mendoza-Denton, & Keltner, 2012). Now consider a painful or physically demanding ritual endurance ritual, as part of initiation into a community. Even assuming that participation is a signal of social quality, it would be hard to tell if it were (i) a fakable signal of one’s commitment to the group, (ii) a hard to fake demonstration of commitment (e.g., leveraging hard-to-fake emotional reactions, or the fact that it’s harder to endure something we resent), or (iii) a way of advertising a different desirable trait, for example that one has the fortitude and physical capacities to be a useful member of that group (as always, these are not exclusive alternatives). Indeed, in any cooperative endeavor, commitment is only one socially desirable trait among many, and by itself it is often not desirable at all (as anyone familiar with committee work or community orchestras can attest to).

So, the good news for RST is that there are many more signaling models available than just “costly signalling”; more models mean a greater diversity of target systems that can be potentially be modeled as evolved signaling systems. The bad news is that these considerations also entail great complexity in application. While signaling theories are unified by their need for an honest signaling model, the available honest signaling models are quite disjunctive; differing (sometimes opaque) in their dynamics, limitations, and explanatory/predictive potential. The use of signaling theory in biology also suggests that multi-trait, multi-modal causal complexes of signaling systems should be common (Hebets et al., 2016; Hebets & Papaj, 2005). And any signaling that helps explain human religion is unlikely to operate in isolation, and would almost certainly form part of (and be shaped by) larger causal complexes including mechanisms from cultural evolution, the cognitive science of religion, and so-forth.

The question then arises: how much detail need RST get into? This depends on how detailed an empirical science it aspires to be, and the aspirations should be realistic. As noted by Hebets and Papaj, after decades of research on the honeybee waggle dance there are still significant questions about how the different parts of the dance encode information and what other contextual factors are important—and this is on a relatively tractable, biologically universal target system. It would be hard for ethnographic analysis of specific human religious traditions as signaling complexes to approach even this level of detail. But in principle there is no reason why some degree of progress could not be made, along the lines of recent work on social networks and signaling in village-based religious communities (Power, 2017). Signaling theory should be applied to an empirically appropriate level of detail, but at least some of its nuances should be taken into account, if moving beyond simple verbal models.
Signaling theory applied

Thus far, I have argued (i) for a core schematic characterization of RST that is neutral between honest signaling models, and (ii) that the choice of honest signaling models can make a significant difference in what a fully specified theory might end up saying. In general, such a theory would have to plausibly identify senders and receivers, any appropriate signal costs and cooperation benefits, and the means by which sender and receiver strategies are encoded and updated in an adaptive manner. In this last section I propose some templates for more fully specifying RST, in the sense of interpreting signaling models and mapping them onto real-world phenomena.

Applications in the early literature

Some of these modeling complexities (and the possibilities they make available) have only been made salient by recent modeling results, so it would be understandable if they had not filtered out to the more applied literature. Indeed, according to one literature survey, the use of signaling models in the human sciences more generally is often problematic, because “detailed requirements of the mathematical models cited by them appear not to be met … it is most often the differential cost requirement that is the problem although it is typically misapplied rather than ignored.” (Grose, 2011). I argue that there is evidence of something similar in RST, beginning with the proposals made by William Irons (Irons, 1996, 2001).

Irons draws heavily on Frank (1988) for his understanding of signaling. Frank’s view is that (some) emotions and emotional displays have the function of strategically binding individuals to courses of action, but in ways that make this commitment transparent to potential strategic partners. Frank writes prior to much of the formal distinction-making discussed here, but one obvious interpretation of his view is that emotional displays are hard-to-fake/index signals (because acting is hard) which provide insights into the sender’s state of mind. Their strategic value comes from reliably signaling when a sender will react “irrationally”, e.g., refusing to back down from a potentially dangerous confrontation or being cooperative but willing to punish defection regardless of personal cost. This information changes the strategic calculus for observers, allowing them to better navigate interactions with the sender without costly conflict or loss of opportunity, and (in theory) allowing stable, more cooperative equilibria to be reached (i.e., evading the mutual defection equilibrium). Irons appears to see religion working this way: “Religion basically is a commitment to behave in certain ways without regard to self-interest” (Irons, 2001, p. 293).

However, Irons’ terminology is ambiguous. Irons (1996) characterizes religious signals as “costly to fake”. In the later paper they are called “hard to fake”, but with the idea that it is signal costs which are doing the work: “For such signals of commitment to be successful they must be hard to fake. Other things being equal, the costlier the signal the less likely it is to be false” (Irons, 2001, p. 298). No mention is made of cost differential.

This is concerning: hard-to-fake and costly-to-fake are not the same thing. The talk of costs suggests the handicap principle, but if the signals are fakable then simply being costly does not guarantee anything. To recap, imagine a costly initiation ritual. If the prospective benefits for the participants (for being accepted into the group) surpass the costs of entry, then every participant able to pay the cost should take part, regardless of their level of commitment, and any adaptive evolutionary mechanisms should reinforce this (as Sosis [2003] recognizes). On the other hand, if cost exceeds benefit then those who pay will do worse than those who walk away from the deal. Neither possibility allows for the evolutionary reinforcement of a signaling system. Alternatively, it might be that high costs make signals increasingly, teeth-grittingly “hard” to fake due to resentment & cognitive dissonance, pre-empting optimal action. But at first pass that sounds like a non-adaptive explanation where cognitive biases end up doing the explaining.
A more coherent way to make sense of this is framed succinctly in (Sterelny, 2012):

when it is honest, the signal itself is cheap, much cheaper than a fake signal, for it requires none of the scarce cognitive resources of top-down attention, control, and self-monitoring. In Frank's picture of the role of the emotions, cost is relevant to success in signaling commitment by increasing the cost of fake signals. But the mechanism does not act via a handicap principle.

On this interpretation, these are hard-to-fake signals with efficacy costs drawing against limited cognitive resources. What counts against the success of low types are high error rates even when resources are invested and potential failure of the charade when resources run out; not the expenditures themselves or what they might otherwise have done with them.

Sufficiently delineating fitness costs from fitness-orthogonal expenditures is a known issue in signaling theory in general (Bruner et al., 2017; Kotiaho, 2001), which has also been explored in the RST context (Murray & Moore, 2009), and is part of the afore-mentioned currency problem (Grose, 2011). On a charitable reading of Irons though (as per Sterelny), there was no intent to imply that signal “costs” are strategic fitness costs, nor to invoke the handicap principle. Indeed, Irons makes no mention of it or of Zahavi. But costly signaling proper (in the sense of the handicap principle) stages an infringement with subsequent writers. For example, in the context of testing Irons’ theory, Sosis writes:

whenever the gains for defection outweigh the costs of cooperation, the only credible commitment signals are those that are “costly-to-fake” (Zahavi & Zahavi, 1997). If commitment signals are not costly-to-fake, they can easily be imitated by free riders. (Sosis, 2000)

This is a confusing passage, firstly because the phrase “costly-to-fake” does not appear in (Zahavi & Zahavi, 1997), but also because of the move from costly signals being hard to fake (all things being equal) in Irons, to costs being necessary in order for signals to be credible. Pragmatically speaking (backgrounding Grafen’s handicap universality), this appears to conflate costly signaling with hard-to-fake signaling. In a later paper Sosis also cites Johnstone (1997), saying: “Costly signaling theory informs us that the costs of a signal are always conditional; they are dependent on the quality of the signaller” (Sosis, 2003, p. 100). However, Johnstone’s discussion of signal cost was only in the context of Zahavi and differential-cost handicap signaling, not hard-to-fake signaling. But this leads Sosis to question the plausibility that (for example) church attendance is less costly for the conventionally virtuous, and to opt instead that behavior is driven by a differential perception of costs by the sincere and the cynical, and perception-driven selection of signaling strategies is subsequently proposed as the driver for the evolution of religious signaling (Bulbulia & Sosis, 2011; Sosis, 2003, 2004, 2006).

If this reading is correct (and assuming cost-talk in the handicap sense) then it looks like a misstep. Internalized beliefs and commitments might adjust perceived payoff and qualitative experience of participation, but objective fitness payoffs are less plausibly impacted. Informational signaling-like behavior without material fitness differences might be driven instead by entrenched biases and beliefs, but there is no guarantee of adaptiveness in the given strategic context. By the lights of the RST schema this is something of a hybrid: allowing adaptive co-evolution of signaling and cooperation (via positive assortment) but with non-adaptive sender-side signaling strategies. It is still recognizable as a variation on the general theme but arguably blurs the lines between RST and fitness-orthogonal views such as CREDs theory. Again, it also forfeits one of the parsimony virtues of a pure signaling theory (operating entirely at the level of fitness and adaptive selection).

Sosis is clearly aware of the distinction between handicap and index signaling in (Sosis, 2006), and by (Bulbulia & Sosis, 2011) the view seemingly appeals to index signals. For example, in a response to Murray and Moore the authors state that “Honest signals differ from other types of communication because honest signals index commitment-properties such that one cannot easily produce the signal absent the commitment. Honest signals are ‘hard to fake’” (Bulbulia & Sosis, 2011, p. 366). However, the nomenclature here is still somewhat ambiguous. In (Sosis, 2006) “hard to fake handicaps” are
contrasted with “impossible to fake indices”, i.e., attaching “hard-to-fake” to the handicap signaling model (incongruent with the biological/formal literature). But the later paper appears to opt for “hard to fake” signaling, with costs consistent with Sterelny’s reading: “The signal is ‘costly’ in the sense that it is hard to fake, without being financially or reproductively costly per se” (Bulbulia & Sosis, 2011, p. 365). Shortly afterwards they emphasize their more general interest: “evolution has scope to target and amplify mechanisms that give rise to the indexical displays. Such … ‘honest signals’ evolve to enable the sort of cooperative assorting necessary to overcome prisoner’s dilemmas and tragedies of the commons” (Bulbulia & Sosis, 2011, pp. 366–367).

The intention to not treat these papers as a single authorial unit for finger-pointing with respect to terminological consistency, but rather to further illustrate the level-of-detail point. It is clear that Bulbulia and Sosis are interested in honest signaling in the broadly evolutionary sense, incentivised by and co-evolving with increased cooperative assortment. Despite emphasizing particular terms and concepts, they have no axe to grind with respect to the two main branches of signaling. And this is entirely appropriate: any kind of signaling system will do for their purposes (see Wilkins, 2018) for another broad usage). But at some level of detail the differences will become significant. For example, it would be a mistake to assume that the same “honest signalling” mechanism might both explain the inflated, arbitrary cost of a religious signal (implying a differential cost–benefit fakable signal) and be explained by a genuinely hard-to-fake, diaphanous connection between commitment and that same signal (which implies an index model). The two branches of signal form are not just abstractions; they imply different realizer mechanisms with different potential explanatory virtues and distinct evolutionary trajectories (especially regarding signal costs). They should not be conflated or treated as fungible.

In addition, these papers illustrate two ways of responding to the differential objectivity challenge. One option was to drop objectivity in favor of RST-borderline handicap signaling with subjective cost differentials. The other was to drop differential costs in favor of non-handicap, diaphanous index signaling based on honest-emotions or other cognitive constraints. But these are also not the only options. I will conclude the paper by outlining several potential templates for RST, based on cognitive constraints, differential costs, and differential benefits.

**Cognitive constraint templates (hard to fake)**

Beginning with hard-to-fake (index/constraint) signals, we have already seen one interpretation (via Sterelny) where dishonest commitment signals become hard to fake: lies weaving too tangled a web for the liar to reliably overcome.

Another is the Frankian notion that emotional/affective responses provide a hard-to-fake, honest window onto the true commitments of religious participants. E.g. intense or demanding religious rituals as evolved ways of eliciting commitment-indexing affective responses; hard-to-fake displays of enthusiasm/engagement from the genuinely committed, or hard-to-avoid giveaways of dissonance/resignation from the cynical or half-hearted. This ticks several boxes. It includes a clear generative link between prosocial commitment and reliable signals. It also looks like a good fit with many religious rituals and public worship practices, whether high emotional intensity to elicit enthusiasm or dissonance, or low intensity to probe for boredom and resentment (in Quaker meeting for worship, fidgeting is predictive).

But there are also caveats that arise from the preceding discussion. Analogous to the honesty/sincerity distinction, the objects of the emotional displays should fit with the RST schema (i.e., they should imply commitment to the community, not just to its religious trappings). And as an index-signaling model, evolutionary pressures on signal efficacy costs predict that signaling would optimise-away from fitness-impacting demands in favor of discomfort, pageantry, and awe. This is less obviously applicable outside the ritual context: it does not explain demands such as tithing or sacrifice of resources, or restrictions on dress, diet, economic activity, and sexual practices. It
also does not explicitly predict signal complexity or cross-cultural diversity, as the emotional responses of the participants constitute the signal, not elaborate ritual tropes themselves. And any scepticism about the reliability, fakability, or stability of moral emotional displays would have an impact here (see e.g., Bandes, 2016; Deem & Ramsey, 2016). Unfakable-signaling interpretations also rely on exogenous constraints, bringing with them the need for a further evolutionary backstory (as with exogenous cognitive biases in non-adaptive, non-signaling explanations).

Such considerations are not knock-down objections, and for signaling interpretations of religious rituals this might be an appropriate template to use, but they highlight gaps which differential cost–benefit mechanisms might instead fill.

**Differential cost templates**

With respect to differential-cost fakable signaling, we already have one option on the table: Sosis’s hybrid model where the evolution of signaling strategies is driven by differential perceptions of cost (rather than by objective cost–benefit), and the signaling (pseudo-)system is bootstrapped by and co-evolves with benefits from cooperative assortment.

But suppose we also let signals and responses be temporally extended or distributed, rather than always associated with discrete events like rituals. For example, (as suggested earlier), a signaling role might be served by persistent markers or adherence to codes of dress or practice. Signaling strategies involving costly signals also need not be strictly “pay as you go”. A comprehensive survey of all the possibilities will not be attempted, but there are potential examples worth delving into.

The first comes with punishment of defection, which some argue should be folded into the notion of signal cost (Fraser, 2011; Murray & Moore, 2009). Suppose that acceptance into a community after signaling commitment now places you at risk of retribution, should you defect from your commitment. Formally speaking, a signal of commitment followed by defection is equivalent to signaling that you’re a high type but being a low type (high types are the ones who don’t defect). This might seem counter-intuitive, but although the order of signal and response is important in modeling sender-receiver reactions, the order of associated cost and benefit is not: cost and benefit only matter for determining overall payoffs. This means that the whole temporally extended scenario can be modeled as a single step in an evolutionary signaling game, in which both high and low types pay the same (if anything) at the time the commitment signal is sent but low-types who signal also pay an extra expected cost via the risk of being caught out.

Importantly, this is a signaling game with differential costs which fully meets Henrich’s differential objectivity challenge. The evolutionary proviso is that the overall behaviors of the community and the individual agent must be governed by strategies which are updated/reinforced in a more-or-less adaptive response with respect to overall payoffs. If so, then those strategies can co-evolve in a signaling system.

A related template is “bridge burning”. Consider the case of group-specific ritual scarification, tattoos, or other permanent markers. One powerful way of promoting group cohesion is parochial altruism (Bowles & Gintis, 2011): treating ingroup members well while treating outgroup members badly (or simply excluding them—in our traditional evolutionary context humans are obligate community-members). In such an environment, an agent who permanently marks themselves as belonging to a certain group (via group-specific markers like tattoos or ritual scarification) has made a powerful statement of commitment, because they have “burnt their bridges” and dramatically increased their own cost of defection (Sterelny, 2012). If we again ignore the temporal ordering of cost and benefit and only look at overall payoffs, then these are persistent, differentially costly signals: inexpensive for the genuinely committed, but dangerous for opportunists likely to leave the community (or fall out with them). But case the community need not have reliable defection-punishment at all, because bad consequences for defection are the result of general outgroup hostility and the signals themselves.

In both templates, punitive cultural practices can form part of differentially costly signaling systems. Technically, neither exactly maps the classic handicap model, as signal cost is vulnerable to
community/receiver response (e.g., in “punishment”, signaling is only costly for low types who are accepted). However, in both cases the crucial free-rider game path (being low, signaling high, treated as a high) is made prohibitively costly. Neither interpretation is entirely free-standing in the idealized, maximally parsimonious sense either, as they rely on (presumably) exogenous cultural practices (punishment, parochial altruism).

**Differential benefit and other templates**

However, there is a relatively simple differential cost–benefit template which straight-forwardly fits the sort of strategic situations we are interested in while avoiding such complications. Imagine two individuals considering initiation into a local community, call them Flaky and Staunch. The community productively combines the labors of its members and fairly distributes those benefits (making it attractive to join), but there are significant investment/transaction costs for absorbing new members. Staunch is enthusiastically attracted to the group and its way of doing things, but Flaky is just looking to take what’s on offer until a better option turns up. For Staunch, there is little prospect of wanting to leave the community or falling afoul of its rules, but not so for Flaky. So, while Staunch can look forward to staying in the community for life (if admitted), Flaky’s relationship with it will have a comparatively short half-life. In summary, the total future benefits of admission would be significantly greater for Staunch compared to Flaky, while for the community Staunch would be a good investment with Flaky less likely to pay off (for similar reasoning with regard to the evolution of guilt, see (O’Connor, 2016; Rosenstock & O’Connor, 2018)).

There is an obvious way to turn this into a viable signaling game: set a fixed, up-front cost for entry that satisfies condition (3), i.e., one that exceeds the benefit that a fly-by-nighter like Flaky would accrue before moving on. With enough time & experience for sender-receiver strategies to adapt, paying such an entry fee can evolve into a reliable signal of commitment.

This differential benefit, “paywall” template performs well with respect to the ideal explanatory virtues of signaling theory. It provides a clear link between commitment level and signaling strategy in equilibrium, since sender commitment directly influences payoffs. Because differential benefits are doing the strategic work, it predicts signals of significant fixed cost but arbitrary form, and no psychological biases, punitive community norms, or other exogenous constraints/pressures are required (in principle). A community-erected paywall is also simple, scalable and evolvable assuming reasonable variation in commitment levels among potential senders: a small barrier to entry will immediately start filtering out the extremely flaky, incentivising the community to increase it up to some optimum level (beyond which profitable recruitment suffers). Coincidentally, the general differential-benefit template is also similar to Ianacconne’s economic analysis of why strict churches are strong (Iannaccone, 1992, 1994), perhaps allowing this view to be positioned as a differential cost–benefit signaling model.

There are other possibilities for temporally reordering signal, cost, and benefit. For a final example, consider a “learnedness investment” template: spending time in one’s early life studying community-specific religious lore, so that authoritative signals of commitment to that community (and hence reliability as a cooperation partner) can be sent later. Although these later signals superficially look like they are hard-to-fake, the entire process can be seen as a long-run investment: pre-paying the costs for signaling commitment to local groups in a way that is far less expensive if you’re actually interested in belonging to just a subset of them. By being sensitive to these pre-paid signals, local communities can again discriminate between the genuinely (parochially) prosocial, and those more likely to look for other options when the going gets tough.

**Applying signaling theory: summary**

All prior caveats about the difficulty of fitting models to real life cases should be kept in mind when considering these more fleshed-out interpretations. Most obviously, taking part in an intense ritual
might also generate commitment (Whitehouse & Lanman, 2014), and so something that seems to fit a Frankian hard-to-fake signal template (and produce broadly similar results) might operate according to another causal mechanism instead (or as well). And the templates discussed here are by no means exhaustive—variations and additions are of course expected. With such caveats in mind, Table 1 summarizes them.

The options here vary along a number of dimensions: signaling models, cost types and distributions, the temporal ordering of signal-cost–benefit, and special requirements for a working RST interpretation. Each of these also vary in terms of their explanatory virtues, as discussed. The main point to bear in mind is that they are not mutually exclusive, nor do they exclude other (non-signaling system) causal mechanisms. They act on different targets and on different timeframes, and would complement each other. If so, multiple signaling mechanisms might plausibly combine ala (Hebets et al., 2016), and feature in larger causal complexes (with cognitive biases, Big-Gods-style mechanisms, etc) that we might recognize as religions (Alcorta & Sosis, 2005).

In any case, at this level of theoretical detail there is at least the potential to identify and empirically test putative religious signaling mechanisms in the field, based on specific predictions flowing from the relevant signaling models.

**Conclusions**

Time then to summarize what has and has not been argued. In the first section of the paper I argued for a payoff-driven, co-evolutionary reading of RST: religious signals are signals of cooperative

<table>
<thead>
<tr>
<th>Template</th>
<th>Signaling model</th>
<th>Context &amp; application</th>
<th>Signal cost/honesty mechanism</th>
<th>Special requirements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cognitive load (Sterelny)</td>
<td>Hard-to-fake/index</td>
<td>One-off/PAYG (dyadic interaction)</td>
<td>Differential efficacy cost (cognitive effort) prohibiting false signal production</td>
<td>• Complex commitment signal conventions  • Relevant human cognitive limitations  • Affectively/physically demanding ritual conventions  • Relevant limitations in sustained affective “acting ability”</td>
</tr>
<tr>
<td>Emotional display (Frank)</td>
<td>Hard-to-fake/index</td>
<td>One-off/PAYG (ritual setting)</td>
<td>Undifferentiated efficacy cost (pain, time, discomfort) prohibiting convincing false signal</td>
<td>• Perception bias correlated with both religiosity &amp; prosociality  • Cooperation benefits must overwhelm actual signal cost in evolutionary dynamics  • Reliable punishment of defectors</td>
</tr>
<tr>
<td>Subjective cost (Sosis)</td>
<td>“Hybrid” fakable/handicap (subjective differential cost)</td>
<td>One-off/PAYG (ritual setting)</td>
<td>Undifferentiated signal costs made subjectively differential, directing behavior as though strategic</td>
<td>• •</td>
</tr>
<tr>
<td>Punishment</td>
<td>Fakable/handicap: differential cost</td>
<td>Extended/pay-later (initiation or maintenance)</td>
<td>Undifferentiated signal with delayed differential strategic cost (punishment on defection)</td>
<td>• Group-specific commitment conventions  • Rejection based on rival commitments  • Group-specific conventions of learnedness &amp; respect  • Significant costs for learnedness investment  • High signal/entry cost</td>
</tr>
<tr>
<td>Bridge-burning</td>
<td>Fakable/handicap: differential cost</td>
<td>Extended/pay-later (persistent initiation markers)</td>
<td>Persistent signal with delayed differential strategic cost (3rd party rejection on defection)</td>
<td>• •</td>
</tr>
<tr>
<td>Learnedness investment</td>
<td>Fakable/handicap: differential cost</td>
<td>Extended/pre-pay (investment for later signaling)</td>
<td>Differential opportunity cost of investment (committed types only need invest once).</td>
<td>• •</td>
</tr>
<tr>
<td>Paywall</td>
<td>Fakable/handicap: differential benefit</td>
<td>Extended/earn-later (entry requirements for cooperative community)</td>
<td>Undifferentiated signal/entry cost, but delayed differential accrual of cooperative benefits</td>
<td>• •</td>
</tr>
</tbody>
</table>
assurance which (as long as they are more honest than not) allow cooperative societies to stabilize and grow. The RST schema is admittedly a narrow one. But (I argue) it isolates an important explanatory natural kind; offering a more fine-grained comparison of different forms of “signalling” explanation. This can carve the literature in a principled manner, and identify the areas of commonality and dispute (though that would take a more detailed survey of the literature than attempted here). In the second section I argued for a Lewisian understanding of signalling systems as the basis for honest signaling, and broadly outlined mechanisms of honesty-biasing and their implications. Finally, I argued for a detailed “pluralist” approach to looking for religious signalling, and outlined several empirically distinct ways in which it might be instantiated and evolve. The overall approach was to take seriously the formal models which give RST its rationale, and develop more detailed application templates in light of them. The reader may balk at the level of detail recommended, but it is useful to at least acknowledge that such levels exist and can (in principle) become scientifically significant.

It should be stressed that this paper has been a theoretically extrapolative exercise. The discussion has largely avoided relevant empirical work, and whether it is of any applied value is an open question. A second open question is: “why religion?” I have been assuming that religion supplies the content for commitment signaling, but signaling models could equally apply to signals with non-religious content. The apparent ubiquity of religion is therefore not fully explained, and would require a more specific account of why religion (traditionally at least) was the gold standard for filling the commitment-signaling role. This is a project for another day.

In conclusion, none of this has been to say that the RST is a good theory or better than its rivals, just that assessing its full potential requires theoretical as well as empirical work. I have aimed to demonstrate what a more systematic approach to this work might look like, and to argue that the prospects for RST are more varied and interesting than they would seem without such theoretical investment. A deeper engagement with the intricacies of relevant signaling models (including formal modeling and simulation results) would also allow a more fruitful development and more precise, testable hypotheses for empirical inquiry. At this level of detail RST would be mature enough for a more definitive assessment of its prospects.

Acknowledgements

My thanks to Kim Sterelny and Justin Bruner for extensive comments and advice on this paper an associated material, to two anonymous referees for greatly improving the manuscript, and for other helpful comments and discussion including from Martin Lang, Stephen Mann, Matt Spike, Pierrick Bourrat, and participants at the ANU School of Philosophy’s workshop on the Evolution of Religion and the Ettore Majorana Foundation’s workshop “Future Directions on the Evolution of Rituals, Beliefs and Religious Minds”.

Disclosure statement

No potential conflict of interest was reported by the author.

Funding

Research in this paper was funded through the Australian Research Council Laureate Fellowship project (FL130100141). This publication was made possible through the support of a grant from the John Templeton Foundation (Grant ID 60811). The views expressed in this publication are those of the author(s) and do not necessarily reflect the views of the John Templeton Foundation or the Australian Research Council.

ORCID

Carl Brusse http://orcid.org/0000-0001-5732-8750
References


