MODELLING RELIGIOUS SIGNALLING

A thesis submitted for the degree of Doctor of Philosophy
at the Australian National University

Carl Joseph Brusse

February 2019

This thesis is solely the work of its author. No part of it has been submitted for any degree, or is currently being submitted for any other degree. To the best of my knowledge, any help received in preparing this thesis, and all sources used, have been duly acknowledged.

for

Caroline, Alida, and Isobel

Acknowledgements

I owe a great debt to my supervisory panel members past and present: Kim Sterelny, Justin Bruner, Rachael Brown, and Ben Fraser, for their deeply insightful commentary and gentle corrective during the various phases of this project. In particular, Kim very much provided both the inspiration for this project and the impetus for venturing to return to philosophy for this PhD candidature. I would not have entered into this without his encouragement, and I certainly would not have made it out the other side without his support and mentorship. Thank you, Kim, for your judgment and generosity, and for your astute, necessity-sensitive application of patience (or lack thereof) when required. Justin too deserves special mention for patiently schooling me in evolutionary game theory and signalling games (despite my sometimes bizarrely misguided intuitions), and in suggesting and contributing to my first simulation projects and our joint paper at the 2016 PSA, published in Philosophy of Science.

Being at the School of Philosophy at the ANU, the number of faculty and grad students (local and visiting) who have taken the time to listen and offer valuable insights would be too long to record here. A few of the locals who deserve special mention are Ste Mann, Chris Lean, Ron Planer, Anton Killin, Heather Browning, David Kalkman, Alex Sandgren, Dominie Dessaix, Donald Nordblom, Ross Pain, and Ten-Herng Lai. I would especially like to acknowledge Matt Spike for the many hours he spent sitting next to me, helping me get my most recent simulations to work. Other philosophers from here and elsewhere whose insights I have mined include Paul Griffiths, Emily Parke, Brett Calcott, Toby Handfield, Russell Grey, Kevin Zollman, Pierrick Bourratt, Colin Klein, Seth Lazar, Geoffrey Brennan, and Alan Hájek. Given the interdisciplinary features of this project, I have also been lucky enough to benefit from the intellectual generosity of many expert scholars in the science of religion, especially Joseph Bulbulia, Aiyana Willard, Eleanor Power, Martin Lang, Quentin Atkinson, John Shaver, and two anonymous reviewers from Religion, Brain and Behavior.

Regarding material support, I must thank my current employer, the University of Sydney, and Professor Paul Griffiths in particular for his generosity, support and understanding over the last months of write-up. Thank you, Paul. The ANU's support to me as student has included travel scholarships, and too many free meals to count as part of the 'thesis boot camp' and veterans' days program. These are the sessions where much of this thesis was written up, along with

funded ‘shut up and write’ sessions organised by PARSA, and informal sessions organised by myself and others, including Adam Bugeja and Hanh Nguyen. The comradery and productivity of the group writing environment was a revelation.

Finally, but most importantly, I want to thank my wife – Caroline Margaret Brusse. Without her encouragement and support I would not have started, continued, or finished this thesis. Her firm conviction that I could get it done, and her shouldering of disproportionate burdens and responsibilities while I flailed my way towards submission is the single most important reason it is now completed (and we can get something of our lives back). This dissertation is dedicated to her, and to our children Alida and Isobel, both of whom were born while it was being written.

Abstract

The origins of human social cooperation confound simple evolutionary explanation. But from Darwin and Durkheim onwards, theorists (anthropologists and sociologists especially) have posited a potential link with another curious and distinctively human social trait that cries out for explanation: religion.

This dissertation explores one contemporary theory of the co-evolution of religion and human social cooperation: the signalling theory of religion, or religious signalling theory (RST). According to the signalling theory, participation in social religion (and its associated rituals and sanctions) acts as an honest signal of one's commitment to a religiously demarcated community and its way of doing things. This signal would allow prosocial individuals to positively assort with one another for mutual advantage, to the exclusion of more exploitative individuals. In effect, the theory offers a way that religion and cooperation might explain one another, but which that stays within an individualist adaptive paradigm.

My approach is not to assess the empirical adequacy of the religious signalling explanation or contrast it with other explanations, but rather to deal with the theory in its own terms – isolating and fleshing out its core commitments, explanatory potential, and limitations. The key to this is acknowledging the internal complexities of signalling theory, with respect to the available models of honest signalling and the extent of their fit (or otherwise) with religion as a target system. The method is to take seriously the findings of formal modelling in animal signalling and other disciplines, and to apply these (and methods from the philosophy of biology more generally) to progressively build up a comprehensive picture of the theory, its inherent strengths and weaknesses.

The first two chapters outline the dual explanatory problems that cooperation and religion present for evolutionary human science, and surveys contemporary approaches toward explaining them. Chapter three articulates an evolutionary conception of the signalling theory, and chapters four to six make the case for a series of requirements, limitations, and principles of application. Chapters seven and eight argue for the value of formal modelling to further flesh out the theory's commitments and potential and describe some simple simulation results which make progress in this regard.

Though the inquiry often problematizes the signalling theory, it also shows that it should not be dismissed outright, and that it makes predictions which are apt for empirical testing.

CONTENTS

1. Cooperation and Religion: evolutionary approaches

This dissertation is about a class of evolutionary explanations, signalling theories of religion, which promise to explain two features of human society that are otherwise surprising when seen through the lens of evolutionary theory. One feature is our high level of social trust and cooperation, even between strangers, and often in the absence of clear and immediate benefits. The other is the prevalence of socialised religion, whereby people submit to seemingly arbitrary demands and restrictions – again, without obvious immediate material reward, and at some cost. Both are unexpected. Even if human beings had been preconfigured to find hypersociality and religion subjectively attractive, the grim logic of game theory and selection at the individual level dictates that these behaviours and these favourable pre-configurations should whither on the vine (or be nipped in the bud). Indeed, our closest great ape relatives exhibit little evidence of such frivolities. But these features appear to be universal and stable in traditional human societies, forming what might be dubbed (by hypothesis) a socio-religious phenotype. The evolution of such a phenotype would demand a more sophisticated naturalistic explanation, and the signalling theory is an explanatory approach that would bind the two phenomena particularly closely; by interpreting religious behaviour as a signal that facilitates cooperation and cohesion within a community.

However, the signalling theory of religion is also burdened by being an interdisciplinary hybrid; and it draws on formal, theoretical, and empirical resources that are diverse and subtle. Disciplinary priorities tend to skew more heavily toward one of those three at the expense of the other two, meaning that some of those subtleties can be easy to miss. For example, formal and theoretical resources with respect to signalling as an evolutionary theory (i.e. in the abstract) have progressed in recent decades, so one overarching goal in what follows is to draw out some lessons from that work and re-apply them to religious signalling theory. The approach I take is a reappraisal of what modern signalling theory has to offer and the development of a conceptual and modelling-based toolkit, tailored to religion as a target of explanation.

The eight chapters of this dissertation are organised into three progressive stages. The first two chapters provide conceptual and empirical backgrounding to position the signalling theory within the literature and with reference to the phenomena it purports to explain. Chapters three to six explore and develop the signalling theory ‘from the inside out’; starting with an abstract characterisation and working through various nuances of signalling theory to explicate the

possible applications of a signalling theory of religion and its associated predictions and explanatory potential. Chapters seven and eight turn to formal modelling; demonstrating the potential of formal analyses for answering questions about what the signalling theory of religion can explain, and for generating testable predictions.

The rest of chapter one is divided into three parts. Because religious signalling theory is characterised by the link between cooperation and religion, both of these need a firm backgrounding. Section 1.1 is a conceptual review of the cooperation problem, to better pin down what sociality means and survey some of the theoretical approaches that have been taken to it. Section 1.2 does the same for what is now generally referred to as the science of religion. In both cases I will be arguing for a way of organising the leading options in the literatures, in a way that will be most relevant and useful for the middle part of the dissertation. In the case of religion for example I will focus on a mechanistic understanding of explanatory theories, understood within an evolutionary framework, and in relation to the relevance of different explanatory mechanisms to the cooperation problem. Section 1.3 then pivots to considers the archaeological evidence with respect to *origins*: what we say about the timing and sequence of social and religious traits in the evolution of humans and their societies. These details speak to the theoretical approaches already laid out, and to the viability of religious signalling as a significant explanatory mechanism. Chapter one therefore seeks to outline and impose some sort of order upon three voluminous bodies of literature. To avoid turning it into long lists of authors and contributions, I err on the side of elaborating distinctions and frameworks, rather than providing comprehensive literature surveys.

One key distinction developed by the end of this chapter will be that between explanatory mechanisms and theories in the full-blown sense. The thesis is primarily about mechanisms. Religious signalling as I will investigate it is primarily a research tradition that identifies a family of explanatory mechanisms with various implementation options and potential utilities. In the second chapter I conclude this backgrounding work by way of contrasting it with a full-blown theory that has been recently developed and defended by several scientists of religion: the Big Gods theory. The Big Gods theory is an impressive explanatory construct, which I deconstruct and critique using the tools from chapter one.

In chapter three I introduce and defend the core mechanistic *role* that I see as characteristic of religious signalling theory: the unifying rationale that allows us to recognise signalling

mechanisms as a unified and scientifically interesting explanatory kind. The characterisation is a functionalist one, and I investigate the implications of that while backgrounding the details of how different signalling mechanisms might fulfil the role. Those details begin to appear in chapter four, which outlines broad distinctions between major families of signalling models – their strengths, limitations, and evolutionary differences. The notion of costly signalling is introduced here with caution, as the notion of costly signalling has a confused and confusing history of usage in the social sciences. This chapter ends with a taxonomy of the families of modelling options for religious signalling theory – index signalling and varieties of differential cost-benefit signalling – and an argument that the differences between these need to be taken seriously in application to real-world phenomena. The investigation of these differences is continued in chapter five, via making explicit the game theoretic formalisations of differential cost-benefit signalling models. Alternative payoff structures significantly impact the likelihood of signalling, and the parameter ranges (such as signal cost, and prior probability of cooperation) over which signalling is an evolutionarily stable strategy. Details matter.

In chapter six we return to consider the application of this work to religious signalling theory. Via discussion of the early literature on religious signalling theory, I show that the distinctions between modelling approaches are often important. The differences highlighted in the previous chapters mean that different signalling models connect to differing strategic profiles: contexts in which we need to distinguish what is at stake for senders and receivers. I therefore develop a number of ‘templates’ for applying signalling models to rituals and other religion-relevant settings – these are suggestions for better connecting models to target systems.

Up to this point, the focus is on mining the existing theory of signalling in general for its application to religious signalling theory. The last two substantive chapters flip the dialectic and pose questions for formal modelling, based on certain particularities of religious signalling, and the previous chapters’ conclusions about apt ways to model them. Chapter seven reviews the current formal literature salient to these questions, identifying gaps and unanswered questions in the literature. This motivates simulation analyses that I outline and report the results of in chapter eight. While these results are preliminary and/or suggestive and point to future work, they demonstrate both the explanatory value of that work and the methods by which it might be carried out.

The overall intent of this dissertation is therefore not to validate or refute religious signalling theory as *the* (or even 'a') explanation of religion, but rather to investigate the nuances of its explanatory potential, and to validate interdisciplinary connection-building between formal, philosophical, and applied approaches to the human sciences. I will be concluding that the modelling options for religious signalling mechanisms are far more varied than generally understood, but that they all come with applicability conditions and limitations that must be respected on a case-by case basis. There is, in other words, a predictable trade-off between generality and explanatory power, but (at least at the theoretical level) the inclusion of signalling mechanisms in co-evolutionary explanations of human cooperation and sociality looks promising.

1.1. Explaining costly behaviour

The central evolutionary puzzle about both sociality (by which I mean a proclivity for cooperative, norm-governed social behaviour) and religiosity is similar: they are costly. They can be costly in the sense of imposing direct demands on individual human beings, such as donations of resources to the group, tithes to a religious body, long stretches of time devoted to community or religious activities (perhaps including aversive experiences), or restrictions on dress, diet, and reproductive freedom. They can also be costly in the sense of being risky: to embrace standards of generosity, trust and loyalty is to put ourselves in a vulnerable position should our conspecifics fail to reciprocate, or otherwise turn on us. Our generosity can be taken advantage of, and our trust and loyalty can be betrayed. Finally, they are also costly by omission and restraint. Respecting cooperative, communal, and sacred values often closes off otherwise advantageous action options, such as appropriating valuable resources or simply looking to one's own interests *prior* to the interests of others or the expectations of religion. These costs, risks, and restrictions can have tangible fitness impacts, especially in traditional, forager or subsistence agrarian societies where there are slim margins between prospering, surviving, and being in real trouble. It would be a mistake to imagine that humans are the only creatures to suffer from strange, costly behaviours, but equally foolish to ignore the special and ubiquitous sophistication and diversity of them in our societies. At least from the naïve evolutionary perspective (as from that of naïve rational egoism), it is remarkable that disapproving terms like 'free rider' or 'unbeliever' ever came to be.

The relevant literature here is vast and ranges across many disciplines: biology, anthropology, economics, psychology, sociology, philosophy, and others. For example, the psychology of how adherence to values and norms becomes internalised, and/or how one's social identity 'fuses' with that of the group, such that these effects can override selfish imperatives will be part of any full explanation. No doubt it will also appeal to historical or contextually particular causes (the evolution of this or that brain region, contingencies of the Pleistocene climate, or in the particulars of the agricultural revolution, etc). As this is a PhD thesis rather than an encyclopaedia, the coverage here will be mercenary. In the spirit of the proximate-ultimate distinction, I will be tackling cooperation from the perspective of evolutionary theories which might help explain the proximate psychological (etc) mechanisms associated with social/religious behaviour. I will be largely taking those 'social preferences' (as (Bowles and Gintis 2011) calls them) as given as features of our evolved cooperative psychology, and will be concentrating on a systematic overview of the relevant theoretical approaches. The initial discussion will therefore be highly abstract, without many particulars or colourful examples of cooperative and/or religious behaviour in humans. Only in section 1.3 will a (highly selective) survey of the human historical/evolutionary evidence be introduced, in order to compare against the theoretical background, and to provide background the discussion of signalling theory in later chapters. First theory, then anomaly.

1.1.1. Framing the cooperation problem

Complex cooperative effort extended over time (i.e. where payoff and reciprocation are not immediate) requires the honest investment by all members of the band who can trust each other to fairly share out resources, workloads, and information. This is the classic problem of investing in cooperative practices. A bunch of indiscriminately trusting Kumbaya-types will always be an attractive target for would-be shirkers and bullies, because being a shirker or bully (if you can get away with it) conveys a fitness advantage. We would expect such populations to be invaded by free-riders, or (as shirking behaviours come in degrees and can be learned or innovated) for free-rider traits to be ratchetted up over time to exploit the incautious extension of trust. This renders indiscriminate trust evolutionarily unstable. With indiscriminate trust alone, unless evolutionary interests are perfectly convergent, complex cooperation should never have got off the ground. This at least is a broad statement of the problem, described in a human context.

Following (W. D. Hamilton 1964a; 1964b) we can define altruism and cooperation in the abstract; to distil the problem, set the relevant terms, and frame the possible responses to it (West, Griffin, and Gardner 2007). Table 1-1 shows a modified version of Hamilton's 2x2 matrix by which a social action can be categorised (a social action being one which has cost/benefit impacts on both the actor and at least one other individual – the recipient(s)). The top left cell classifies actions which benefit both the actor and her interaction partner(s). The cell below that includes actions which benefit the interaction partner(s) but are detrimental to the actor. Both classes of action might be labelled 'cooperation' but the later type of cooperation is 'altruism' in the technical sense; the conferring of benefit on others at personal cost to oneself, which cries out for a better explanation. Both cells on the top row are equally unproblematic: mutualist cooperation benefits the actor the same as exploitative behaviour[1]. The remaining cell is 'spite': acting in a way that harms the recipient while also at a cost to oneself. Genuinely spiteful behaviours, if commonly observed and not the result of error, are just as problematic for first-order evolutionary explanation as altruism.

Table 1-1 **Categorisation of social behaviours according to payoff valence, positive (benefit) or negative (cost), for actor and recipient(s) (after (Hamilton 1964a)).**

		Recipient(s)	
		Positive payoff	*Negative payoff*
Actor	*Positive payoff*	cooperative (mutualism)	exploitative
	Negative payoff	cooperative (altruism)	spiteful

It is therefore the bottom row of the table which poses the *prima face* problem for evolutionary and/or game-theoretic explanations of human social behaviour, because both altruism and spite appear to exist in human societies. Especially in forager societies, we pool or share food and other resources, and act according to norms of generosity and self-restraint (the opportunity costs from refraining to exploit are effectively a form of altruism). We also often enforce our internalised norms via third-party punishment when minding our own business would be more

[1] As well as active predation, exploitation, betrayal, and so-forth, this category also technically includes passive exploitation, where the actor 'acts' to accept the benefits of the *other* agent's altruism.

profitable, i.e. by going out of our way to punish or berate the guilty at our own expense, perhaps risking backlash or reprisal. Because such behaviour may promote cooperation, it is sometimes called 'altruistic punishment' (Robert Boyd et al. 2003). The relative importance of these two types of behaviours (and the relationship between them) is a matter of active debate[2], but for our purposes we can treat them as of the same, explanatorily problematic, kind for a fitness-based account of evolution.

With the problem stated this way, there are several ways of approaching the problem and explaining these phenomena. Most broadly, we can either look elsewhere for explanations (i.e. outside of fitness-based evolutionary explanations) or seek to re-frame the cited real-world phenomena so that altruism and spite behaviours collapse into mutualism and exploitation.

1.1.2. Inclusive fitness and kin selection

One potential re-framing strategy is via Hamilton's notions of inclusive fitness and/or neighbour-modulated fitness (W. D. Hamilton 1964a; 1964b)[3]. A trivial case of altruism which conforms to this explanation is parenting. On the face of it, parenting involves the dramatic surrender of a considerable share of one's resources and future potential, in order to create new human beings. But of course, this is naïve: psychological altruism and evolutionary altruism are not the same thing (Sober and Wilson 1999; Stich 2016) and maximising biological fitness is not co-extensive with this sort of crude rational egoism. Fitness-maximising actions are those which maximise one's chances of having the most descendants, and so securing the survival and onward fitness of one's own children is not altruism at all from evolutionary perspective, it is an act of evolutionary self-interest. Even a transfer of resources to my children which lowers my 'narrow' personal future fitness to some degree (in terms of having further children) but improves theirs to a much greater degree, can be a net benefit to my overall fitness.

Inclusive fitness generalises this idea of an overall fitness to include other blood relatives: I can improve my inclusive fitness by paying a personal fitness cost in order to improve the fitness of close relatives, not just direct descendants. The condition for such an action being

---

[2] See for example (K. Jensen 2010), and a series of papers by Patrick Forber and Rory Smead on the classification of social behaviour, and how spiteful behaviour can lead to fairness (Forber and Smead 2014a; 2014b; 2015; 2016).

[3] As with most writers, I will concentrate on inclusive fitness rather than neighbour-modulated fitness. For a clear contemporary comparison of the two, see (Birch 2016).

'fitness-rational' is the famous Hamilton's rule inequality $rb > c$, i.e. the benefit b to the recipient multiplied by the fractional relatedness r of the recipient to the actor must exceed the personal cost c to the actor[4]. Thus, Haldane's supposed quip[5] that he would sacrifice himself to save the lives of eight cousins (r = 0.125) or two brothers (r = 0.5). If an ostensive costly donation to one's kin satisfies this condition, then (in terms of inclusive fitness) it is actually mutualism, not altruistic cooperation. Inclusive fitness becomes evolutionarily significant if an altruistic behaviour is more common among kin due to its heritability, and kin (and the trait) are positively assorted for some reason. Under these conditions – when the altruist trait benefits others who share it – selection can favour the evolution of that trait. This is kin selection.

Inclusive fitness and kin selection are popular tools in biology which many biologists will defend as providing great insight (Abbot et al. 2011), but it is not clear how explanatory they are in the case of the human societies we are interested in. As we shall see, while forager societies often place great emphasis on being able to trace kinship connections, they are also often divided up into highly mixed interaction groups, making questionable a primary reliance on kin selection. For ostensive altruism to evolve by kin selection, there must be something that biases the positive assortment of relatives. If an actor is no more likely to interact with a close relative than a phylogenetic stranger, then there is no inclusive fitness incentive for cooperative behaviour. The two possibilities to secure positive assortment are i) viscously structured populations where offspring do not widely disperse, so are considerably more likely to interact with parents and siblings than they would in a mixed population, or ii) kin recognition: the ability to recognise (via observable traits) who is kin and who is not, with a preference to interact (cooperatively) with them. Studies of kin and non-kin interactions in

---

[4] A more descriptively general form of the equation might be $\sum r_b b > \sum r_c c$. I.e. the sum of beneficiary relatedness-times-benefit should exceed the sum of the donor relatedness-times-cost, considering every cost-paying donor and benefit-absorbing recipient (both types in relation to the agent whose action is being assessed). This allows calculation of actions which impose costs on others as well as the actor, and which can benefit more than one recipient. In the spirit of Haldane-meets-trolley problems, it would then be fitness-rational to throw no more than eight cousins onto the tracks to save two brothers, if no other means were available. The more familiar formulation is the special case of an act where there is a single beneficiary and the only donor is the actor (who is self-related at r = 1).

Note also that Hamilton's rule also makes sense of spiteful actions: relatedness can be negative if the recipient is less related to the actor than the population average (represented by r = 0), predicting that it might pay to expend one's own resources on damaging the fitness of non-kin competitors (especially in a smaller population where a greater relative advantage is obtained).

[5] See (Birch 2017, 1–4) for an entertaining overview of the mythos surrounding Haldane, Hamilton, and the early days of inclusive fitness theory.

humans support the idea of closer cooperation with kin to some degree, but the evidence is mixed and the significance of Hamilton's rule, inclusive fitness, and kin selection is disputed (Nowak, Tarnita, and Wilson 2010; Birch and Okasha 2015). And in larger societies neither assortment mechanism is adequate to explain the common practice of cooperation with non-kin.

1.1.3. Strategies and reciprocity

Another re-framing strategy is via diachronic, contextual, 'holistic' or, 'life-strategy' reframing of the problem cases (call this the holistic strategy). On this view, instead of individual acts of generosity, punishment, or norm-following, it is human capacities and dispositions for these behaviours (over an extended time) that are the proper units of selection. On this longer-term view, a community of unconditional co-operators look less like altruists in the technical sense. The benefits conferred on others at a cost on a short timescale are reciprocated by others following the strategy, meaning that when viewed on a whole-life timescale, their unconditional cooperation strategy is mutualism not altruism. But this is a naïve redescription and the basic problem remains: the first member of that kumbaya community to turn exploitative would receive a much greater payoff and therefore be evolutionarily favoured. A population following unconditional cooperation strategies can be invaded by agents following unconditional 'exploit' strategies and is therefore unstable. The natural next step is to consider conditional strategies and reciprocity.

But these sorts of strategic considerations call for a more sophisticated, game-theoretic formalism. Table 1-1 characterised the cooperation problem in terms of how to characterise the actions of a single agent, in relation to (passive) others. The standard way to represent the cooperation problem as a strategic situation is the prisoner's dilemma, represented in another 2x2 matrix table 1-2. Both actors (players) have symmetric options: they can 'cooperate' with or 'defect' on the other, and the numbers x, y in each outcome cell represent the payoffs for players 1 and 2, respectively. Note that the two tables do not map on to one another, because the rows/columns now denote actions rather than resultant costs and benefits. However, if actor 2's action is held fixed, then the actor 1's choice is a choice between interactions that match Hamilton's classifications (formally, if not entirely intuitively). Cooperating with cooperation is mutualism (1, 1), cooperating with defection is altruism (-2, 2), defecting on cooperation is exploitation (2, -2), defecting on defection is (in effect) spite (-1, -1). But the specific payoff values here mean that these now have a determinate preference ranking: exploiting > mutualism

> spite > altruism[6]. That is to say, a player benefits most by defecting on a co-operator, with mutual cooperation being second most valuable, mutual defection third, and being defected on while cooperating the least preferred. This ordering means that the only stable combination of strategies is mutual defection, since it is better to defect on both cooperation and defection. Originated with game theorists in the RAND corporation in the 1950s with the name coined by Albert Tucker (Tucker 1983; Poundstone 1993), the prisoner's dilemma has become the standard formal representation of the cooperation problem and linked (accurately or otherwise) to social dilemmas such as the tragedy of the commons (Hardin 1968).

Table 1-2 Payoff matrix for the Prisoner's Dilemma: standard form game for two actors

		Actor 2	
		Cooperate	*Defect*
Actor 1	*Cooperate*	1 , 1	-2 , 2
	Defect	2 , -2	-1 , -1

A well-established literature exists with respect conditional cooperation strategies in the cooperation problem; originating (for our purposes at least) with Trivers' model of reciprocal altruism (Trivers 1971) and Axelrod's demonstration of the relative viability of a 'tit-for-tat' strategy in simulations of repeated interactions (Axelrod 1980a; 1980b). Axelrod's computerised tournaments pitted different strategies against each other over many iterations of a prisoner's dilemma game. In a single iteration of the prisoner's dilemma game, the best move is to defect (as no matter what the other player's move is, defecting delivers the superior payoff). But in the iterated prisoner's dilemma a different dynamic is introduced which can permit a cooperative equilibrium path (previously known via the so-called 'folk theorem'). Tit-for-tat is a strategy where a player begins by cooperating, but in subsequent rounds conditionalizes on the behaviour it received in the round prior. Tit-for-tat does marginally worse than 'always defect' in head to head matchups (as it will be exploited in the first round). But if

[6] Note that the number values here have been chosen to emphasise the similarity to Hamilton's categorisation table. The more common notation is to use four non-negative numbers for the four possible payoffs types, but what is important is the ordinal ranking of their rational preference.

two tit-for-tat players meet then they will do as well out of each other as two ‘always cooperate’ players would – while the ‘always cooperate’ players would be brutally exploited by ‘always defect’. Surprisingly, in open competition between various strategies (some of them far more complex), it is hard to beat tit for tat in terms of overall performance. And, though there are differences here that are important in other contexts, the notion of reciprocal altruism is broadly analogous: a strategy which is sophisticated enough to recognise cooperation and reward it with further cooperation can do better under certain conditions.

It is possible then that in a competition of many potential strategy types *over an extended timeframe*, conditional cooperation strategies might do better than constant defection/exploitation, even in extremely competitive strategic situation approximating the prisoner’s dilemma. It is even tempting to see this as a model for the evolution of behaviour-guiding cooperative norms and social preferences: conditional strategies that were constructed out of various social attitudes and dispositions because they paid off for their adherents over some relevant evolutionary timeframe.

But there are well-known formal details here which throw spanners in the works (see (Bowles and Gintis 2011) for a more thorough overview). First, tit-for-tat and similar strategies are not robust to error, because if one of a pair of tit-for-tatters mistakes the other’s action for defection then they will collapse to mutual defection. Second, in a finite sequence of games there is an incentive to defect in the last round of the game, but if such behaviour is available as a strategy then this incentivises even earlier defection (you want to the first to defect, not the last), collapsing the cooperative equilibrium. In other words, cooperation is stabilised by *some* ways that players might conditionalize behaviour on the game context, but other ways (beneficial to the player) will destabilise it. Games of more than two players (n-player games) are unstable for similar reasons, as any defection by some players can likewise trigger a cascade of further defection (you don’t want to be the last one cooperating).

Conditionalization in the form of punishing defectors might incentivise cooperation. But unless punishment comes from outside the game like a bolt from the blue (which is cheating by changing the payoff structure so it is no longer a prisoner’s dilemma), meaningful punishment must involve some players harming others, either exploitatively or spitefully. The former looks very like defection, the later raises a second-order cooperation problem: by what method is the costly punishment stabilised? Either way, simple forms of punishment such as going on strike

(i.e. withholding cooperation or productive labour in a collective project) should just undermine cooperation, especially in n-player games. Such tactics can work in real human societies because they can shame the offenders back to good behaviour. But shame as a robust moral emotion (or more generally, a normative social disposition) is akin to the propensity to punish defectors by being exactly the sort of proximate mechanism we are trying to explain as conditional strategies. Strategies need to be incentivised somehow to be explicable in evolutionary terms. This is not to say that punishment strategies are a dead end, and they are the subject of ongoing formal and empirical research, e.g. (Robert Boyd, Gintis, and Bowles 2010; Bowles et al. 2012). But such studies show that will have to be highly specific and targeted (especially in the n-player game) and preferably coordinated, requiring a high degree of sophistication and supporting social infrastructure. This suggests that punishment may play a greater role in stabilising mature cooperative societies, rather than bootstrapping them into existence.

These are some of the standard problems with reciprocity and conditional strategies as explanations of cooperative societies, some of which will reappear in application to the human evolutionary timeline in section 1.3. Real-world conditional strategies that support cooperation would have to be quite sophisticated and multi-layered, as well as more forgiving than would be ideal for short term payoff maximisation, making their robustness against exploitative invasion less certain.

There is one final methodological worry to flag before moving on, concerning the general shift of focus from actions to strategies. The worry is that by focusing on strategies we are perhaps inappropriately idealising and over-rigidifying the actors that these strategic models are supposed to represent. Strategies are the basic ontological units for game-theoretic formal analysis – agents adopt and exchange strategies, but strategies are generally not modelled as being constructed out of anything else. Modelling idealisations generally ignore the messy psychological reality of how human behavioural strategies are complex consequences of more casually basal phenomenal: memories, ideals, hungers, loyalties, emotional affect etc. Human beings are not ideal agents. The normative or behavioural strategies that human beings follow in social contexts can be expected to be malleable and/or fragile, and often opportunistically

applied, with an unpredictable connection between endorsed principles and actual behaviour[7]. So, unless unsophisticated social strategies are conceived of loosely enough to be realised by basic temperament (i.e. a probabilistic disposition for annoyance and retaliation), it is dubious to suppose that they are highly robust, developmentally or genetically entrenched selected effects. We should not be too precious here with respect to demanding all the details, but the applicability of game theoretic strategies to human behavioural strategies is an open question.

A tentative corollary of this is that the more powerfully specific the supposed strategy, the less plausible it is to project its evolution onto real populations. Arbitrarily complex strategies can be codified and vertically transmitted generation-to-generation (like alleles) only in computer simulations. Treating strategies as basic units of selection is a simplifying idealisation that should be kept in mind before we move from studying the dynamics of games themselves (an interesting academic pursuit in its own right) to drawing explanatory lessons for human social evolution. And a stronger potential concern can be pressed as well: while we can study idealised models of precise, neatly defined strategies subject to selection pressures, it is questionable whether humans really have consistent and robustly action-guiding strategies, as opposed to ostensive principles on one hand, and patterns of behaviour on the other. So, it is even more questionable that they might all at once be sufficiently powerfully specific, sufficiently stable over time, and sufficiently heritable to be subject to selection. Addressing such worries is by no means impossible, and the literature on the psychology and evolution of norms might be of great relevance here[8]. But there is a thicket of questions with regard to how formal strategic models might be best applied to human reality.

1.1.4. Group selection

For our purposes, it will help to think of inclusive fitness and strategic/reciprocity as 'reframing' approaches. Each alters the accounting system for cost and benefit and shifts the

---

[7] See for example the literature on the classic 'good Samaritan' experiments (Darley and Batson 1973). Mutation rates are often included in the update dynamics in models of strategy evolution to account for imperfect conservation of strategies over time. However these tend to be quantitative: e.g. there being a chance of an alteration in the agent's likelihood of cooperating or not conditional on some relevant variable (see (Gintis, Smith, and Bowles 2001) for one good example of this). The point here is that strategy conservation in real agents should also depend in a qualitative way on the psychological and epistemic dependencies of that strategy, with some being more fragile than others, with mutation having a 'collapse' bias.

[8] Especially on the psychology of norms e.g. (Sripada and Stich 2006), and on the nature and potential evolutionary role of externalised moral norms, such as (Sarkissian 2016), and (Stanford 2018) with associated commentaries e.g. (J. P. Bruner 2018; C. J. Brusse and Sterelny 2018; T. Davis and Kelly 2018; Stich 2018).

operational focus of evolution away from the agent's immediate ostensive actions and payoffs. For kin selection, the focus is extended to include the payoffs of relatives. Reciprocity/conditionalization shifts it from acts to strategies and/or extends the payoff time horizon. Each approach has potential, but this comes with complications and limitations especially in the real-world context.

A similar move to acknowledge (in the broad respect of reframing the first-order problem) is to appeal to group or multi-level selection theory, which has it that altruistic individuals proliferate not due to their individual fitness, but due to the fitness of the group they are a part of relative to competing groups. Fitness accrues at the group level, either directly distributed among its members to turn losses into gains (so-called multi-level selection 1, or MLS-1 (Heisler and Damuth 1987)), or alternately driving the propagation of groups as higher-level individuals (MLS-2). If anything, group selection is far more complex and controversial than kin selection and reciprocal altruism. For example, MLS-1 is often seen an occasionally-useful heuristic but mathematically equivalent to standard fitness considerations (Kerr and Godfrey-Smith 2002), or better folded into a more general conception of inclusive fitness and social evolution (Birch 2017). MLS-2 is a more radical proposal, but problematically so in that it posits adaptation at the group level, not just selection (Gardner and Grafen 2009), and a degree of group unity over time that appears implausible, at least in the human case (see 1.3). For either version to work, there must also be i) significant inter-group competition, and ii) robust disincentive against group members who are insufficiently altruistic, perhaps via problematically costly punishment (Williams 1966)[9]. Recent work on multi-level selection has also focused on the conditions for it in major evolutionary transitions such as the evolution of multicellular life (Okasha 2006; Clarke 2014), with some theorists following on from original suggestions in (Szathmáry and Smith 1995) that the evolution of distinctively human forms of sociality constitute a more recent transition (Powers, Schaik, and Lehmann 2016). How literally to take this characterisation is controversial.

Group selection in general has also seen more conventional use in a number of explanatory projects with respect to human sociality, notably (Boehm 1999), (Wilson 2003), and (Bowles and Gintis 2011). Wilson for example argues that altruistic punishment can be an "altruism

[9] Indeed, it is with Williams' influential objections to early group selection that the cost-shifting/circularity of altruistic punishment was first highlighted.

amplifier" by being cheap with respect to costs paid vs costs inflicted; though it is far from clear how this should be modelled (Okasha 2003). It also appears to assume an asymmetry in retaliation capability (as an equal ability to punish the punishers will tip the scales back again). Bowles and Gintis instead make a strong case via formal modelling for the viability of a complex group selection process to explain social preferences. More specifically, their view is that group selection drove the evolution of a form of parochial altruism: altruistic behaviours directed at in-group members, and spite-like behaviours directed at outsiders. The idea is that biological and cultural traits for in-group cooperation co-evolved as the flipside of outgroup hostility; in a competitive group selection process where the more cohesive groups wiped out the less cohesive groups (and their members). Cohesiveness was thereby selected for despite the individual costs involved in looking toward group welfare rather one's own. The reason that I will not be considering this view further is that it seemingly requires a level of competition and conflict between ancient human forager groups that is empirically suspect (Sterelny 2016). Group selection remains an intriguing possibility, especially if inter-group competition over cohesion might be reimagined in a more indirect manner[10], but it is one that I will be largely leaving to one side.

A related debate I will merely acknowledge is that about the independence or otherwise of group selection from kin selection and reciprocal altruism. As already noted, (Birch 2017) argues for a framework (based on Hamilton's rule) that would encompass MLS1 group and kin selection, and the *formal* equivalence of these (under certain assumptions) is well understood[11]. In contrast, (Sober and Wilson 1999) describe reciprocal altruism and kin selection as special cases of group selection, though this arguably relies on an unjustifiably broad conception of groups (M. Barrett and Godfrey-Smith 2002). As Barrett and Godfrey-Smith note, each mechanism works by engineering the correlation of behaviours: by kin preference/limited

[10] For example, (Robert Boyd and Richerson 2009) also identify imitation and selective migration as potent mechanisms of cultural group selection, and stress that even violent conflicts can be reasonably bloodless, with losers merely intimidated into retreat or assimilated by the winners. Another mechanism might be differential vulnerability of small-world economies to collapse under environmental stress; e.g. when drought strikes, or the game dries up. For obligate co-operators such as humans in forager societies, having one's group disintegrate under such conditions (or just fail to function efficiently) is bleak prospect. Such relative intrinsic robustness might be just as effective as relative capacity for warfare in generating differential prospects for groups, and incentivising looking after one's own.

[11] Though see also Birch's defence of the applicability conditions for group vs kin selection descriptions of social behaviours.

dispersal, group structured populations, or strategy conditionalization. This is a key point: the positive assortment of co-operators is crucial for the endogenous[12] evolutionary explanation of cooperation. Whether there are deeper connections is not a subject for now.

1.1.5. Strategic context

To conclude this section on responses to the cooperation problem we can consider one final reframing strategy: to reframe the cooperation problem itself. In the discussion of the conditional strategies I introduced the prisoner's dilemma as a strategic model of the cooperation problem, and pointed out that imagining ad hoc, exogenous punishment of defectors counts as 'cheating' by altering the stipulated model. Likewise, it is cleating to take as given any other evolutionarily improbable cooperation-stabilising mechanisms, like complex social institutions or predilection for altruistic/spiteful punishment. But it is not cheating to question whether the prisoner's dilemma is the most appropriate model for the predominant strategic contexts of ancestral human environments.

Table 1-3: Comparison of three payoff matrices for modelling strategic social interaction

(a) Prisoner's dilemma

	C	D
C	1 , 1	-2 , 2
D	2 , -2	-1 , -1

(b) Stag hunt

	C	D
C	2 , 2	-2 , 1
D	1 , -2	-1 , -1

(c) Coordination

	C	D
C	1 , 1	-1 , -1
D	-1 , -1	1 , 1

Table 1-3 reproduces the prisoner's dilemma in abbreviated form (1-3a) and contrasts it with the two other common 2-player games, the Stag Hunt (b) and a coordination game (c). The row player and column player have the same moves available, labelled C and D. In the coordination game there is no conflict of interest and the players prefer the same outcomes, where their behaviours complement each other. There are two equilibria in this game: both play C, and both play D. The common illustration is choosing which side of the road to pass on: it doesn't matter which side is chosen, as long as both players make the same choice. The stag hunt shares

[12] Exogenous explanations might include actual punishment of defectors by God, or selective breeding by human-farmers unknown. Less facetiously, the distinction here is between evolutionary explanations that can be understood solely in terms of the population in question (evolving a 'normal' environment), and those that require some sort of special pleading in terms of 'quirky' environmental forces that are themselves unexplained in an evolutionary sense.

features of both the coordination game and the prisoner's dilemma[13]. As in the prisoner's dilemma, it makes sense to interpret C and D as 'cooperate' and 'defect' because defecting is always costly for the other agent. But unlike the prisoner's dilemma, defecting only makes things better for the player if the other player defects: the rational response to cooperation is cooperation. Like the cooperation game then, the stag hunt has two stable equilibrium end-states C-C and D-D, but with the cooperative (mutualism) equilibrium being more valuable. The stag hunt is the game which the prisoner's dilemma turns into if exploitation is significantly punished, or in some other way moved one place down in the preference ranking of outcomes.

Brian Skyrms has argued that the stag hunt has been unfairly overlooked in the formal literature as a potential model for cooperation problems (Skyrms 2003). First, prisoner's dilemma situations can be transformed into stag hunts in a number of ways (other than directly modifying the payoff structure), including by the iterated play of prisoner's dilemma games under certain conditions[14]. Because it is rational to cooperate when you expect your partner to cooperate, the robustness of the cooperative equilibrium requires *trust* to win out over risk-aversion. Skyrms demonstrates several formal results, for example: in interactions between strangers we should not expect the cooperative equilibrium to do significantly better than the defection equilibrium, but in interactions between neighbours, cooperation is much more likely (eventually). Behavioural 'nudges' that we can reasonably imagine being delivered by simple social institutions (i.e. that upregulate trust) could therefore make a significant difference in collapsing populations toward cooperation and be strongly incentivised for that.

So, while escaping the prisoner's dilemma is usually seen as the gold standard for cooperative explanations, it would make a real difference if this turned out to be setting the bar too high with respect to the actual evolutionary history of human cooperation. We know that human beings did learn to cooperate, and (assuming the learning was a fitness-driven evolutionary process) it is reasonable to suppose that they did so unaided by freak exogenous intervention. If so, then perhaps part of the way out of the *human* cooperation puzzle (within the adaptive

---

[13] The term stag hunt comes from an example story by Rousseau: two hunters each have the choice between working in concert with the other to hunt a stag, or 'going it alone' to catch a hare (meaning the other is left with nothing, if they embark on the stag hunt alone).

[14] For example, the threat of retaliation by a vindictive partner (aka the 'shadow of the future') can change lower the value of immediate defection in the prisoner's dilemma, by lowering the expected payoff from all future iterations.

evolutionary paradigm) is to re-evaluate the problem we are trying to solve; by looking in more detail at the mapping between strategic models and the human evolutionary past.

The question then becomes about the strategic environments our evolutionarily-recent human ancestors lived within – with respect to our cooperative social preferences (however these are specified) and how they might have evolved in response. There too many gaps and moving parts here to allow for much more than speculation, but we can make a few observations and gesture at a potential argument. To prefigure section 1.3 somewhat, an interesting feature of the prisoner's dilemma is that the examples associated with it tend to be rather 'Holocene' in character, compared to the hunting interpretation of the stag hunt. E.g. the tragedy of the commons is a problem because moving on to greener pastures (as foragers do) is not an option. Even the eponymous interpretation of the prisoner's dilemma (two prisoners facing the choice whether or not to rat each other out to the authorities) implies layers of norms, hierarchies, and other social institutions that might be quite alien to human enculturated within Pleistocene social worlds.

This suggests a possibility: perhaps prisoner dilemmas were simply not as common in the past as they are now. The prisoner's dilemma is the hardest strategic environments for cooperation-promoting mechanisms to stabilise because it involves the highest level of social entanglement and mutual vulnerability. No doubt there are many 'naturally occurring' prisoner's dilemmas to be found in loose, unstructured social worlds, for example with allocation of limited resources such as meat from a valuable kill. But their number is demonstrably increased once there are (for example) obligate common resource conventions to be exploited, or institutional power to which to betray criminal conspiracies. To have evolving norms, social institutions, economic lifeways and so-forth is to have evolving strategies, but also an evolving game. Increased cooperation and new ways of cooperative life become possible, some of which might introduce new strategic challenges (especially at their fringes). Therefore, perhaps then the evolution of cooperative capacities was less a response to prisoner's dilemmas than the evolution of prisoner dilemmas was a response to cooperative capacities which began their upward trajectory in earlier, less entangled times. A co-evolutionary theory of cooperative strategies and strategic environments would have to be informed by archaeological evidence and ethnography to substantiate such game theoretic speculations (this will be briefly returned to in section 1.3).

Any such empirically fleshed-out account would need to have a narrative skeleton roughly as follows. First, conventions naturally arise in coordination games of mutual interest. These conventions breed a certain degree of trust, which bleed over into stag hunt-style situations to beneficial effect. Further conventions and assurance devices would be beneficial in this sort of social environment, such as socially shaming the risk averse as cowards as a weak form of punishment, in an environment where it is *beneficial* to evolve a sensitivity to shame[15]. Such social infrastructure in turn rewards closer, more economically entangled lives – potentially addressing the 'other' cooperation problem of generating benefit (see Calcott 2008). This makes prisoner's dilemmas more common, which incentivises more stringent social infrastructure, and so-on, up to a point where institutional mechanisms are sufficient to stabilise cooperation in an environment where prisoner dilemmas predominate.

Key to any such approach would be a move already discussed: appealing to the blurring effects of evolutionary time horizons, lack of information, or lack of strategic specificity. The grain at which selection acts will therefore be crucial. On one hand, if selection were highly specific, then it may not matter much whether prisoner dilemmas or stag hunts were more common because distinct stag-hunt strategies would not 'bleed over' into cooperation in prisoner's dilemmas. On the other hand (metaphorically speaking), if evolution does not know which games you are going to play and the contexts in which you play them, it is unlikely to be able to optimise your strategies to each class of games you actually face in the moment. The best it can do is provide some kind of satisficing toolkit. Likewise, agents unsure whether they are in a stag hunt or a prisoner's dilemma (or unable to switch strategies on so rapid a timescale) will tend to evolve strategies which hedge their bets and do acceptably well in the prevailing strategic mix of that environment.

While this explanatory possibility is highly speculative, it demonstrates the need to carefully consider the mapping between evolutionary models and the actual evolutionary facts as we know them.

---

[15] As, unlike in the prisoner's dilemma, there is a more valuable cooperative equilibrium in the stag hunt which it can be beneficial for the risk-averse to be nudged towards.

1.2. The problem of religion: maladaptation and cultural evolution

The previous section was entirely focused on sociality and the traditional cooperation problem, neglecting religiosity as a distinct problem. To a certain degree, explanations of the costliness of religion will mirror many of the same moves we have just seen. Acts of religious devotion often include donations to the religious community: barn-raising in Amish communities, tithing, or other transfers of resources from the individual to the community (with religious authorities taking a cut), or from individual to individual. But many other costly forms of religious expression are not just costly to the actor, but also yield no discernible benefit for anyone at all. Giving up one's foreskin to the community is not a donation (at least not one of any significant material value). Yet circumcision is a relatively widespread religious rite in religious societies both developed and traditional, and extreme forms of ritual, non-therapeutic mutilation (with attendant risks of infection or long-term impairment) are not uncommon. Stunning, widely-discussed examples of this include the male initiation rites traditionally practiced by many Aboriginal Australian peoples involving penile subincision: cutting open the underside of the penis to convert the urethra into an open channel (Pounder 1983). More prosaically, simply attending a long religious ritual is an opportunity cost with respect to how that time might have otherwise been spent, but these are loss-making exercises with no material upside to anyone. Likewise, acquiescing to seemingly arbitrary restrictions on dress, diet, and behaviour removes options that might be potentially valuable (especially when subsistence is marginal) but pay no communal fitness dividend.

The actual costs or benefits (overt and hidden) of individual examples can be disputed, of course. Circumcision may be protective against disease. Donations toward functionally useless religious infrastructure might be economically and culturally stimulating, with indirect benefits. Food taboos might have roots in sensible hygiene practices. But it is hard to deny that much religious activity seems straightforwardly wasteful, irrespective of how the payoffs are framed.

This breaks with the 2x2 matrix of social behaviours that we began with in the previous section. Neither altruistic nor spiteful, much costly religious activity is simply useless. And the costs of useless actions cannot be explained by mechanisms like inclusive fitness, reciprocity, or group selection, which rely on there being *some* benefit to redirect or re-interpret. This is arguably the real mystery of religion: the widespread, normativised, and very public (apparent) waste of resources and potential that it encourages and demands. Perhaps because of this, many

scientific approaches to understanding religion propose maladaptive explanations, some of which dovetail with maladaptive explanations of cooperation. I turn to these after a quick acknowledgement of an elephant in the room: the fact that what constitutes religion is disputed.

1.2.1. Religion's demarcation problem

Religion is a many-faceted phenomenon, and it tends to get described by researchers in ways that say as much about their own research questions as they do about the phenomena. That is certainly the case for early attempts at pinning it down. For example, for William James, religion was: "The feelings, acts, and experiences of individual men in their solitude, so far as they apprehend themselves to stand in relation to whatever they may consider the divine" (James 1902). Emile Durkheim famous defined religion as "a unified system of beliefs and practices relative to sacred things, that is to say things set apart and forbidden - beliefs and practises which unite into one single moral community called a church, all those who adhere to them." (Durkheim 1912). Note the difference in focus; even if these authors were not famous founding figures in their modern academic fields, it should be obvious that Durkheim was a sociologist and James a psychologist. The classes of target systems which each give the name 'religion' (the psychological phenomenon and the social phenomenon) reflect the academic context, methodological toolkit, and objects of study which each are familiar with. Wield a hammer and every problem looks like a nail.

There is of course some commonality between these two outlooks, in the form of an ascription (by believers) of sacredness or divinity. Indeed, an earlier foundational thinker, anthropologist Edward Tylor in his seminal *Primitive Culture* used a working definition of religion as belief in 'spiritual beings' (Tylor 1871). Religious beliefs and religious practices are most obviously identified as such because of their apparent orientation toward the spiritual or supernatural.

But generalising from this is difficult. Notions of the supernatural, of spiritual beings, sacredness, and divinity are slippery and contested, and using them to precisify an extension of the concept of religion (i.e. demarcating between beliefs/practices that are religious and those that are not) leaves some uncomfortable descriptive mismatches. For example, if 'sacredness' is spelled out in a practical/behavioural sense, to mean special or unquestionable reverence, then that might over-generate to describe intense patriotism as religious (as just one example). With regard to divinity or supernatural separateness, problematic cases that challenge a natural-supernatural distinction include the kami of Shinto, which are in some sense

‘concealed’ but also part of the natural world (Gall 1999). Some religions such as Jainism vehemently oppose the notion of powerful divinities, portraying what ‘supernatural’ entities as exist in their cosmologies as mortal and more akin to humans than the Abrahamic notion of God. If we loosen the notion of ‘spiritual’ entities to include such entities, and others such as souls or karma, then a ‘supernaturalist’ definition along the lines of Tylor’s would be more inclusive. But there is a danger in this case that ‘spiritual’ or ‘supernatural’ just comes to mean something not attested to by modern science, inviting comparison to Hempel’s dilemma and other well-worn philosophical problems with defining physicalism and the physical (Ney 2008; Stoljar 2010). Just in terms of mismatch with common sense, this would also be potentially problematic because the ether and phlogiston were clearly not religious concepts, but nor (arguably) are beliefs in psychics, UFOs or the occult. Religious belief-systems might be loosely described as ‘bad science’, to the extent that they try to explain the world as it appears by positing concealed forces and/or entities, but they are more than *just* bad science. Or at least that’s the presumption of a positive demarcation project for religion regarding supernaturally flavoured descriptive content. The ultimate ‘test’ question for this approach is whether ‘naturalistic religion’ makes sense, or whether putative cases of it (which might include Confucianism, some forms of Jainism or Buddhism, and some contemporary ‘Christian’ sects such as non-theistic Quakerism) are simply descriptive confusions.

An alternative and perhaps more ecumenical approach is to instead look to *normative* content or patterns of commitment as definitive. One contemporary theologian for example describes religious faith as “a comprehensive worldview or ‘metaphysical moral vision’ that is accepted as binding because it is held to be in itself basically true and just even if all dimensions of it cannot be either fully confirmed or refuted” (Stackhouse 2007). But again, this approach threatens to leave awkward gaps and over-extensions, and makes more sense as an extrapolation from Abrahamic religions then as a genuinely cross-cultural definition. As even Tylor noted, traditional religions often encode and engage with moral beliefs and practices to only a minimal degree, even when the moral beliefs and practices of their adherents are sophisticated and well-developed. Environmentalism, Marxism, libertarianism, and other moral/political systems all represent potential problem cases in the other direction. It is hard to give a definitive knock-down argument against the possibility of a happy medium here, but scepticism is justified.

As noted by (Dawes and Maclaurin 2013) among others, there is very little commonality in content between all the things which get described as religions. Perhaps not surprisingly, many scientists of religion downplay the importance of belief to religion, and defend the view that ritual and other pragmatic patterns of behaviour are more important (Sosis and Kiper 2014) as well as (not coincidentally) more stable and empirically tractable. For example, (Rappaport 1999) argues that ritual is a) constitutive of religion, b) socially fundamental in terms of relating people to one another, and c) in many cases responsible for *creating* the sacred, not just responding to it (e.g. by sanctifying objects, people, etc). On this approach, religion is foundational to traditional human social life, and the proper study of religion and its social significance is captured by the study of such rituals, big and small. It is tempting to think that a modern science of religion as ritual might simply pin down the subject matter by surveying it thoroughly. I.e., look at all the forms taken by "religious life" (broadly construed), and let a more precise grasp of what we're talking about just fall out of the commonalities and family resemblances which are discovered. But it we should be equally sceptical that such a change in focus will find many characteristic universals across (for example) Roman Catholicism, Jain ritual, foragers religions, and cargo cults. Religion is weird in the sense that it is both pervasive and incredibly diverse. Like language, all human cultures seem to have it, but it's hard to find any common core between them[16]. It is not clear if there is anything unifying at all about all the things we might want to call religion, other than our disposition to call them by that name, and certainly it is difficult to come up with a set of necessary and sufficient conditions which picks out all and only the things we tend to want to hang the name on.

Several fall-back approaches are possible of course. One is to bite some arbitrary bullets: Jainism is not a religion, Marxism is, etc. Another is to give up on 'religion' and instead study beliefs in deities, or sacred values, as their own separate, operationalizable domains of inquiry. A more hopeful option is to give up on necessary and jointly sufficient conditions for religion, and make do with 'fuzzy' characterisation rather than demarcation (Sosis 2009). Indeed, this is entirely compatible with studying religion as a social phenomenon ala Durkheim, but some sort of principled choices will need to be made with respect to inclusions, at least to pick out representative case studies. A popular version of this is a 'dimensional' approach: to specify a of indicative features, such that all undisputed cases of religion have *some* of those features to

[16] For the sake of the analogy, I side with the opponents of Universal Grammar.

some degree. An influential example of this can be seen in (Smart 1989), however the dimensions cited here (e.g. practical/ritual, mythic/narrative, experiential/emotional) are so broad that they (by the author's own admission) apply just as well to *any* powerful world-view: religious or secular. At the very least, such approaches are obliged to allow (or ignore) borderline cases of religion and are resigned to a degree of ambiguity about the limits of the concept. Another fall-back is to give up on the term religion altogether and focus on one or more of the indicative traits as interesting targets of scientific study themselves. In this way, study of (for example) religious ritual and its role in created 'identity fusion' with a group (Whitehouse and Lanman 2014) can be continuous with the study of ritual and fusion in the context of warfare (Whitehouse 2018a) or football hooliganism (Xygalatas 2018).

So it is doubtful that religion (as a scientifically interesting concept) is satisfactorily captured by focusing just on a certain phenomenon[17], as religion and religiosity are more than any of these individually. But outright eliminativism about religion is impractical, at least for current purposes. It would be tempting to cheat and stipulate a working definition along the lines that religions are those parts of cultures toward which resources are uselessly expended (though this of course would face its own over-generation problem). But this is not required. In evaluating a theory like the signalling theory of religion and the role it might have played in the evolution of human society, it is not strictly necessary to have a cut and dried description of religion. Siding with Sosis and others, it can be good enough that we can recognise it in paradigm cases and tolerate the existence of borderline cases.

But we can (and should) say more than this, because there are two ways of being 'borderline'. First, there can be borderline cases by vagueness: e.g. how highly a cultural practice or belief system scores along a set of characteristic features/dimensions (supernaturalism, ritual, etc). But there is also ambiguity: exactly which dimensions should be included in that characterising set? A paradigm religion like Roman Catholicism will score highly on all plausible sets of characteristics and so be recognised as such. But it may be being so-recognised for different reasons; and focusing on the evolution of different feature-set configurations (and traits that enable them) might still lead in different directions. This means that we still need empirical reality to be such that there is an actual 'cluster kind' (whatever the cluster might be) for us to

[17] Which (to recap) might include cognitive capacities, social predilections, belief contents, propensities to believe, or indeed a culturally evolved piece of social infrastructure.

be latching onto. This broadly coheres with the Quinean notion of ‘family resemblance’. I.e. we can identify religious practices as such not by exhibiting necessary and sufficient traits/conditions from some list of religious traits, but because subsets of such traits tend to go together. The justification of ‘religion’ as a concept comes because every culture exhibits a substantial subset of these traits (despite no core, essential feature that they all share) at a level above what might be expected if they were just random cultural tropes, and (crucially) because these traditions are more similar to each other than they are to other cross-culturally recognisable traditions (such as foraging practices or athletic traditions). If so, something being ‘religious’ would depend on intrinsic features to some degree but mostly on extrinsic context: the fact that religious feature subsets reliably recur and persist, and in a way that ‘stick out’ as being salient and distinctive. Whether this means that ‘religion’ constitutes a natural kind is a question I won’t address.

Other than these metaphysical and empirical assumptions, there will be other limitations on scope with respect to investigating religious signalling theory. For the most part, I will be talking about religions as *public* phenomena – belief is important, but only insofar as it influences participation in religious practice and the integration of that practice into social life. So, rituals, behavioural norms, and associated social infrastructure will be the focus. But the assumption will be that certain ‘supernatural’ beliefs are especially apt for playing a facilitating role. I.e. while I follow (Bloch 2008) and others in seeing religion as deeply integrated into social life, I don’t assume that is all there is to say about it. Part of what tends to make religious practices distinctive and potent is that they are packaged up with cosmologies which are used to justify the practices but are themselves epistemically intractable, in that they include ‘exotic’ entities and forces outside of evidence and everyday experience. It is not just that the ritual is done, it is that it is accompanied by a story (believed by all or not) about why the ritual is important; a story which reaches beyond the mundane. This is the (admittedly hand-wavy) notion of religion and religiosity that I will be proceeding with.

1.2.2. Evolutionary explanations of religion: adaptive, mismatch, and by-product

We can now position the signalling theory in comparison with other approaches in the contemporary science of religion. However, the demarcation problem for religion is mirrored by a similar diversity with regard to these approaches; both in terms of what aspect of religion they focus on, and the types of explanations that are offered. Correspondingly, there are several

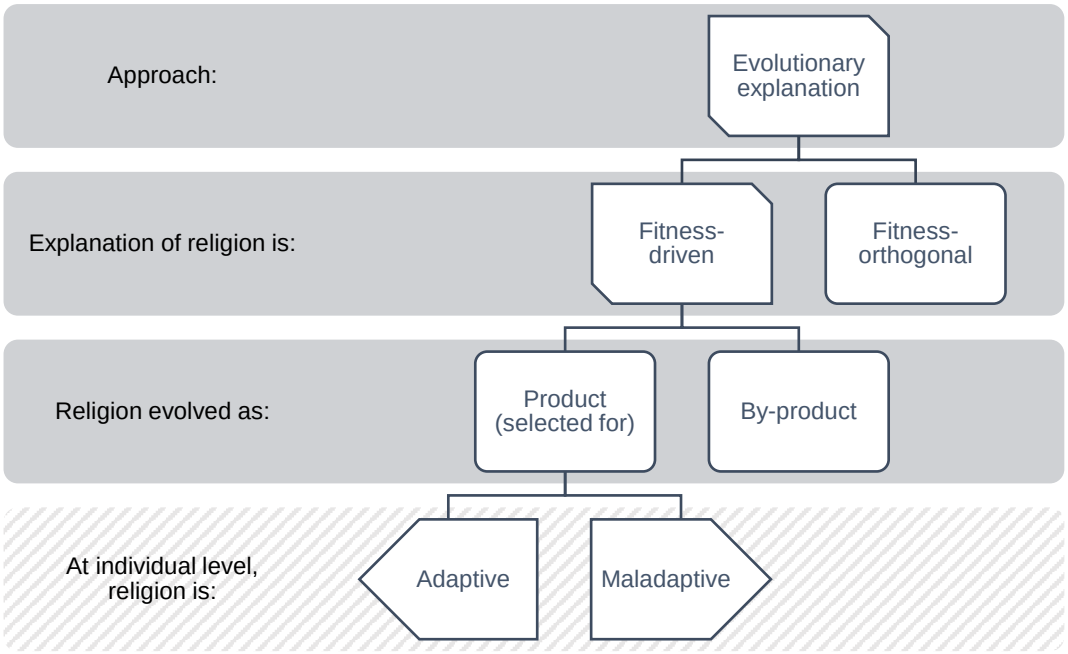
ways of carving up the space of evolutionary explanations, all of which require a certain amount of shoehorning with respect to borderline cases and hybrids.

As summarised by (Sosis and Kiper 2014), one initial way of splitting the literature is into two main disciplinary streams: the approaches of cognitive scientists, and the those of researchers who use the tools of behavioural ecology. According to Sosis and Kiper, evolutionary cognitive science approaches tend to see the proclivity toward religious beliefs, commitments, and behaviours as the un-selected by-products of other cognitive traits that were positively selected for. Classic expositions of such approaches can be found in (Atran 2002) and (Boyer 2001). Atran for example frames the preface of his book with the question "stone age minds for a space age world?", and surveys possibilities around the idea that religion is a cognitive by-product of other cognitive systems (by-product views are common enough that (Powell and Clarke 2012) call this the 'standard model'). Behavioural ecology instead focuses on religious behaviour and how its costs and benefits can be made sense of in an economic or evolutionary sense. While this distinction loosely reflects the focal aspect difference between James and Durkheim (i.e. between religion as a psychological or social phenomenon), it is possible for the disciplinary and focal distinctions to cross-cut. We can have behavioural ecology explanations of beliefs via practices, or cognitive science approaches to the forms that religious practices take.

It makes most sense in the current context to carve up the literature in relation to the cooperation problem and its evolutionary explanations. For example, (Bourrat 2015) divides evolutionary explanations of religion into by-product explanations (including (Boyer 2001)), individual fitness explanations, and group-level fitness explanations, discussing the latter two in the context of religion evolving as a solution to the cooperation problem. Bourrat also points out a truism which will be central to later chapters: different types of explanations are not mutually exclusive, especially if they involve selection on different aspects of religion or over different timescales. Theorists might be combining different approaches in a single overall model, so it is important to discriminate between and compare the potential components model of overall models. As this thesis is focused on one particular model component (signalling, as defined in chapter 3) and the actual review of the literature will be kept to a minimum, this is the sort of taxonomy I will favour here.

We can break down the possibilities according to categorical answers to several questions; as shown in figure 1-1. Staying agnostic about the primary religious phenomena to explain, evolutionary approaches will either explain them as the result of fitness-driven adaptations, or as fitness-orthogonal processes. Examples of the latter include drift and (some forms of) cultural evolution. Examples of the first type are natural selection and sexual selection at the individual level, or kin selection or multi-level selection. The next question (for fitness-driven explanations) is whether the religious traits were directly selected-for, and are therefore the products of these processes, or are just by-products of other selected traits.

Figure 1-1: A conceptual breakdown of evolutionary explanations for religion



I will treat the question of whether the religious traits are individually adaptive in their current/observed/relevant[18] environment is one that cross-cuts with the other distinctions, rather than representing a further branch of classification with respect to origin explanation. Mismatch explains why a trait is costly, but not why it is. It can be due to selection having acted at a level other than the individual, to environmental change since the trait was selected for. Contingent environmental change can also generate a mismatch between fitness-produced

[18] Given the context of enquiry, 'current environment' might usefully be extended back to something more distant but ethnographically proximate; ignoring post-industrial modernisation or looking at 'traditional' societies to screen out the effects of significant cross-cultural contamination.

traits and their environment, or a match for by-products and orthogonally evolved traits (though this is less likely). In principle then there are six possible classifications for religion; three (pure) evolutionary origin types – product, by-product, and orthogonal – which might be adaptive or maladaptive at the individual level (evolutionary products also decompose by mechanism of selection). This is my preferred way of organising the terms in the literature, such as adaptation/adaptive (sometimes used co-extensionally), by-product, and maladaptive (sometimes used for both maladaptive products, and orthogonally generated traits).

Here then are a few of the main explanatory themes in the literature, according to this schema:

Adaptive product

- **Signalling theory**: religious ritual is a costly/honest signal that helps solve the cooperation problem by improving positive assortment (see chapter 3 onward).

Maladaptive product

- **Group selection**: religion (broadly speaking) is a group-level adaptation that improves group fitness by helping solve the cooperation problem (Wilson 2003; Haidt 2013).
- **Social technology**: religious ritual evolved via cultural group selection to induce cohesion/binding with the group and its priorities (Whitehouse 1995; Whitehouse and Lanman 2014).

Maladaptive by-product

- Hyperactive agency detection disorder (**HAAD**): there is an otherwise fitness-enhancing cognitive bias toward the attribution of agency (since false positives are less dangerous than false negatives). Religious beliefs are by-products of this over-sensitivity (Guthrie 1995; J. L. Barrett 2000; Atran 2002).
- **Precaution systems**: ritualised behaviour is produced by a (generally) fitness-improving cognitive/neurological system which evolved to trigger/automate precautionary behaviours. Religious rituals (and their justificatory constructions) are by-products of this (Boyer 2001; Boyer and Liénard 2006).
- **Existential crisis**: as a by-product of expanding cognitive capacities, humans came to fear their own mortality and seek comfort or meaning not forthcoming from the natural world (Freud 1927; Bergson 1932).

Maladaptive orthogonal

- Numerous cultural-evolutionary explanations (see 1.2.3)

Some of these explanatory approaches have already been discussed in passing (and others will be raised in what follows) and some interpretation is also required in mapping them to my categories. But a comprehensive overview of them is not vital to the research project in any case. More important is the relative positioning of signalling theory, which is one of the few theories to uncontrovertibly posit an individual fitness benefit for religious behaviour. One other notable claim to individual fitness benefits has been made by Dominic Johnson, who argues that the self-restraint from 'bad' behaviour caused by fear of supernatural punishment was actually individually adaptive (Johnson 2015; Johnson and Bering 2006; Johnson and Krüger 2004). But this explanation appeals (in part) to changes in the human strategic context such as the emergence of language, arguably a game-changer with regards to the ability of cheaters to avoid detection and spiteful punishment. Johnson argues that fear of divine punishment helped avoid such traps, i.e. HAAD-style cognitive biases spawned false beliefs, but these turned out to be adaptive given independent contingent changes in the state of ancestral human behavioural ecology. Therefore, although Johnson himself describes the view as relying on individual fitness (and a proper re-interpretation would require more detailed attention) it is at least arguable that it might better be classified as a rare 'adaptive by-product' theory under my schema. Kim Sterelny in (Sterelny 2017b) offers a less ambiguous and more sophisticated evolutionary narrative based on individual fitness, where a series of interactions between fitness, behavioural ecology, and human idiosyncrasies propel the phased 'ratchetting up' of religious and social activity (more about this in section 1.3).

Note that many other views in the literature will straddle the categories given here, by virtue of helping themselves to more than one explanatory process. For example, if the 'existential crisis' explanation is pursued, it might be argued that religion (by salving harmful anxiety) is also an adaptive response. More explicitly, (Fischer and Xygalatas 2014) propose a model whereby extreme rituals both upregulate an individual's devotion to the group (as with the 'social technology' proposals of Harvey Whitehouse and co-authors), and improve the individual's standing among the community (which is a signalling effect). The first effect potentially evolves via cultural group selection, the second improves individual fitness. Another more complicated example of this is the 'Big Gods' theory, which will be used as a case study in chapter 2. Briefly, this theory derives from versions of the 'fear of supernatural punishment' view which (in comparison to Johnson) see cultural group selection as driving the evolution of big-god beliefs, because they improve group cohesion (either directly or by

altering perceived payoffs for punishment (Shariff and Norenzayan 2007; Schloss and Murray 2011)). The Big Gods view expands on this via cultural mechanisms operating on the belief content, so that we now have supernatural beliefs as a) by-products of an evolved observer-bias, but b) triggered/propagated by fitness-orthogonal mechanisms, and c) generating group-level benefits. A prior, similarly composite (but less unified) approach is laid out in (Atran and Henrich 2010).

1.2.3. Fitness-orthogonal explanations: Cultural evolution and innateness[19]

The remaining explanatory category to consider includes fitness-orthogonal processes, which further increase the complexity of the explanatory landscape. The most interesting of these are cultural evolutionary processes, which govern transmission and uptake of cultural components of the human phenotype (beliefs, social behaviours, and perhaps institutions/infrastructure) via cultural learning. These traits are therefore inherited culturally rather than biologically, meaning they can be transmitted 'horizontally' with respect to the germline, between peers and from mentors, rather than only 'vertically' (i.e. from parents). However, not all cultural evolutionary processes are fitness-orthogonal, and it is also important to distinguish notions of selection and heredity in cultural evolution.

We can begin with one example of a paradigmatically fitness-orthogonal process of cultural evolution, Joseph Henrich's proposal that cultural traits (and especially religion) can be propagated via 'credibility enhancing displays' or CREDs (Henrich 2009). A CRED is an action (usually costly, in some sense) performed by an agent, the performance of which would be *surprising* if it were not in line with their otherwise professed beliefs and desires. For example, a performer of the Kavadi, a ritual performed by a handful of devout Hindus during annual festivals, pierces their skin with hooks and skewers over several hours in an act of devotion. It is hard to argue that the performer might be insincere in their religious convictions, because it is hard to see any other motivation to do this. This eliminates one possible reason not to trust them (it's a deliberate con) and leaves only the possibility that they are misguided. Henrich sees religious rituals as CREDs which helped to propagate or culturally entrench their associated beliefs, behaviours, and norms; via social learning and conformity. Importantly, this

[19] Some of the material in this sub-section has been previously published in (C. Brusse 2017). See that publication for a more in-depth discussion of Lewens' views and concerns regarding his conceptual framework.

propagation happens because of the appeal of sincerity rather than any success or payoff-sensitivity. In other words, CREDs are 'maladaptive fitness-orthogonal' cultural-evolutionary mechanisms, which can propagate costly religious behaviour despite the negative fitness impact[20].

But this is not the only template for putative cultural evolutionary processes, and there are multiple potential mechanisms of social learning including 'follow success' strategies (Laland 2004), and there is some evidence of human biases toward such strategies (Wilks, Collier-Baker, and Nielsen 2015). The question is whether such strategies constitute and/or approximate selection processes, and how they figure in the general 'stew' of cultural evolutionary mechanisms.

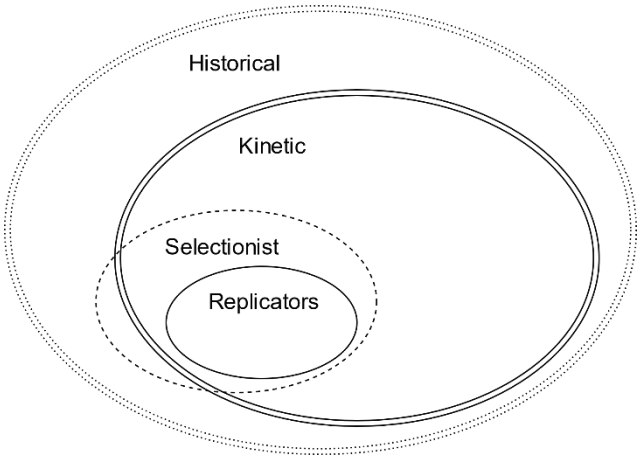
Different theorists will favour different conceptual frameworks to make sense of how fitness-orthogonal and selectionist mechanisms might combine. Peter Richerson and Robert Boyd for example discriminate three kinds of evolutionary process: random processes (mutation, cultural and genetic drift), decision-making processes (guided variation and various forms of biased transmission), and 'natural selection' (the effect that having one cultural variant rather than another has on cultural change within a population) (P. J. Richerson and Boyd 2005). A recent, 'ecumenical' overview of cultural evolutionary theory that also addresses this question can be found in (Lewens 2015). All cultural evolutionists believe that culture evolves, in the extremely weak sense that it changes over time; Lewens calls this 'historical' cultural evolution. Within this category is what he calls the 'kinetic' category, a form of population thinking (in analogy to the kinetic theory of gasses) where cultural change is explained or modelled bottom-up, idealised as occurring via the internal dynamics of populations of more-or-less interchangeable individuals (generally of only a few types). Lewens argues that kinetic approaches are the most fruitful, and he identifies this as the most suitable category to describe the 'Californian' and 'Parisian'[21] schools of cultural evolution, the former associated with Richerson, Boyd, and Henrich, and the latter with Dan Sperber and Oliver Morin. The kinetic approach is contrasted against the more restrictive 'selectionist' approach, which involves an

[20] Note that this isn't straight-forwardly a meme-replicating mechanism because the costly behaviour of the following generations might be quite different from that of their models; e.g. if novel and ever-more extreme devotional displays of are sought out and innovated.

[21] Though these terms instead come from (Sterelny 2017a).

explicit dependence on *fitness*, generally within a populational paradigm (i.e. kinetic), but not necessarily. A subset of kinetic selectionist approaches rely on there being discrete cultural replicators or 'memes' in analogy to genes (Richard Dawkins 1976). A kinetic approach can appeal to social learning and learning biases, and the results may or may not fall under a selection-like description (i.e. involving units of selection being reproduced in a countable, fitness-apt way, akin to genes or organisms). Figure 1-2 illustrates this in the form of a Venn diagram.

Figure 1-2: Venn diagram of Lewens' four classifications of cultural evolution theories



This is not to say that a 'kinetic' theorist will never use fitness or selection-like models, but one advantage of not doing so is avoiding having to talk about the fitness of *what*, and the attendant worries about operationalising the identification and book-keeping of cultural traits which may be ephemeral, dispersed, combine in unusual ways, or admit of degrees. Dan Sperber has argued convincingly against the appropriateness of cultural replicator models, as cultural traits are co-opted and recreated rather than copied and reproduced (Sperber 1996; 2000). Indeed, with a few exceptions such as some forms of artefacts which can serve as discrete templates for copying (Sterelny 2006), the pre-industrial world has few plausible cultural analogues to the allele. But in the absence of a replicator, individuating cultural variants is problematic, as is identifying cultural 'parenthood'; cultural influences also combine in ways far less predictable than sexual reproduction. These problems of individuation (and others) mean that we cannot simply port over models of selection from population genetics, with an inherently discrete basic ontology and inheritance system. Consider a charismatic religious zealot who can either indoctrinate a small number of further zealots, or a larger number of less committed converts: how do we compare the two strategies as 'reproductive', for fitness-

purposes? In other words, should we treat the zealotry and the religion as the same trait (admitting of degrees) or as distinct traits (with a vague boundary)? What if the converts absorb only part of the teacher's ideas or combine them to innovate an overall belief-system radically different from that of the original model? Selection of something implies differential fitness of something, which implies countable descendants of that thing, with reproduction distinguishable from its *growth* and carried out once rather than by repetition, and none of this is unproblematic in the case of cultural traits (Ramsey and De Block 2017).

However, avoiding overt selection of cultural traits does not carve a sharp line between cultural evolution and selection. To begin with, even if selectionism and the notion of cultural fitness can't be made to do any useful explanatory work for the evolution of cultural traits, it is still possible that 'kinetic' cultural evolutionary mechanisms might somehow *approximate* the results of selection in many circumstances, especially where significant costs are involved and sensitivity to success influences cultural transmission. Leaping up a few levels of description, one significant addition to cultural evolutionary theory is cultural group selection (Henrich 2004; P. Richerson et al. 2016). While the details are not important here, the idea is intuitive: common beliefs and other cultural traits might be used to define groups, but may also influence how successful those groups are, and successful groups can better propagate their common beliefs. Different versions of cultural group selection are possible, depending on cultural transmission mechanisms within the group, rates and modes of conversion/migration between groups, how groups are defined (interaction groups, confessional communities within broader societies, etc), how they compete and/or reproduce themselves, and so-forth. At the very least, cultural learning has the potential to empower group selection by increasing the variation between groups (Sterelny 2016).

But there is also potential for intermediate-levels of selection closer to the individual; e.g. Ramsey and De Block's attempt to rehabilitate the notion of cultural fitness relativized to cultural variants possessed by organisms (Ramsey and De Block 2017). A more straight-forward (though radical) way of doing this can be seen in Cecilia Heyes' theory of cognitive gadgets (Heyes 2018). A psychologist specialising in the evolution of cognition, Heyes stands out by focusing on cultural inheritance and the question of how deeply implicated it is in the construction of human minds. Heyes' answer is 'very deep'. Central to her argument is a metaphorical distinction between products of cognition like ideas and behaviours (the 'grist') and the mechanisms of cognition which produce them (the 'mills'). Heyes argues that even

theorists versed in cultural evolution tend to see the mind's mills as biologically fixed points, around which to understand cultural evolution's influence on the grist. This is the assumption Heyes wants to challenge. For her, the mills too are fair game for explanation via cultural rather than genetic inheritance, with only very broad traits (specified in chapter three) being genetically entrenched by natural selection. In this vein, she argues that the very mechanisms for cultural learning that make humans distinctive (social learning, imitation, mindreading, and language capacity) are themselves largely culturally inherited constructs. These cognitive gadgets are non-metaphorical pieces of cognitive technology embodied in human brains, laid down in development by enculturation operating on a minimal 'starter kit' of genetically inherited traits, and enabling other culturally transmitted gadgets to take root.

This notion of cultural evolution therefore combines strong cultural *inheritance* with a more organism-focused (or rather, phenotype-focused) notion of *selection*, since the main influence of cognitive gadget-style cultural evolution is in development. It is also intriguing in this context because it might allow much of the relevant cognitive science of religion (cognitive biases such as HADD for example) to be folded into a unified cultural-evolutionary mechanism along with more shallowly integrated norms and belief contents, which would be equally part of a cultural inheritance but transmitted in ways less reminiscent of natural selection.

Of course, this is a radical proposal. One way of reading Heyes's position is as the intellectual equivalent of an engineering challenge. The human mind is certainly not a pure 'tabula rasa' before enculturation, but how far back might we push it (given high-end estimates of what cultural inheritance can deliver) and still expect to get a recognisable human mind as a result? Heyes makes a strong case for the possibility of only a minimal explanatory role for genetic inheritance, with cultural selection taking the lion's share. But a reality exactly as extreme as the one Heyes advocates would be surprising. It seems plausible that the truth lies somewhere between her view and the conventional hard distinction between biological and cultural evolution, which implies that the appropriate notions of inheritance and selection with respect to various forms of religious traits (beliefs, behaviours, dispositions, etc) are up for grabs.

So, even this quick survey (by no means comprehensive) shows that cultural evolution includes a potential plethora of explanatory mechanisms for religiosity; some fitness-orthogonal and some closer to fitness-sensitive selection, and at potentially at multiple levels of selection. There is an embarrassment of riches when it comes to candidate evolutionary mechanisms.

1.2.4. Fitness traps and syncretic explanation

One additional explanatory approach that fits less neatly into my earlier schema is the fitness trap approach of (Sterelny 2007), which appeals to both fitness-driven and cultural evolutionary processes. Following (Gould and Lewontin 1978), Sterelny describes a fitness trap as "a situation created by a strategy that sweeps the population because it is individually advantageous when not universal, but once fixed, the strategy reduces the absolute fitness of everyone" (Sterelny 2007, 320). With such a broad interpretation (and a generalisation of individuals to agents), this might include familiar 'race to the bottom' phenomena modelled by collapse to a mutual-defect equilibrium in the prisoner's dilemma, and the tragedy of the commons. In international relations, protectionism and militarisation (e.g. nuclear proliferation) would count as familiar model examples: early adopters gain advantages over their competition, and this drives universal adoption and retention of such strategies (because it is in no nation's interest to unilaterally open their borders and/or disarm), but a world bristling with high tariff barriers and nuclear weapons leaves all nations worse off.

However, Sterelny is interested in fitness traps which go beyond this picture and rely more explicitly on frequency dependence, where conflicts of interest within societies drove its members into the arms of strange, seemingly maladaptive cultural traits. His explanation here is also syncretic, because he sees cultural evolution as generating and proliferating maladaptive traits, which then get driven to fixation by a more straight-forwardly selectionist process. His examples including toxic gender relations, dangerous medical and mutilation practices (such as female genital mutilation or foot-binding) and religion. The key, as he describes it, is response-dependant fitness – where the adaptive environment of a cultural trait is in part constituted by the responses of others. In the example of foot-binding among Chinese women, perhaps the response to foot-binding in the marriage market was positive (perhaps because it adhered to an established aesthetic sense or sense of sexual propriety), so the first families to practice foot-binding could 'marry up' the daughters they inflicted it on. On top of any horizontal/learning-based proliferation, there might also be an overall fitness advantage here, and foot-binding (if children robustly reproduce their parent's cultural phenotype) would have headed toward fixation, at least within the aristocratic segment of the population where marriageability was the main venue where women could find a fitness advantage. But any absolute fitness-gains for the early adopters (good marriages making up for the pain and loss of function) will have disappeared as fixation was approached, and all that was left was the

relative fitness penalty for those who rejected the practice. Hence foot-binding is a fitness trap, where it became obligatory for finding a good marriage, and the women in question ended up significantly worse off than their pre-foot-binding ancestors. A similar story might be told in the case of female genital mutilation, or perhaps with military machismo, where fitness-favouring responses might include sexual selection or general improvement in social standing. In the case of religion, the story might be that specific forms of public religiosity initially set individuals apart as potentially generous, virtuous social partners, but this function was lost once these religious practices became universal.

Intriguingly, Sterelny links this to Robert Frank's idea of signal-sending becoming obligatory (Frank 1988), an idea which will be returned to later in the thesis. For the moment though, we can see this an interesting example of how fitness-orthogonal and selectionist processes might can be combined in an explanatory model.

1.2.5. Summary: religion and its mechanisms

In this section I argued for both a cluster-concept of religion (making a few assumptions about what should be part of that cluster) and a mechanistic approach to breaking down the theories that might explain it. The scientific utility of religion as a cluster concept depends on empirical adequacy. Given some of the discussed causal mechanisms though, it seems natural to also see it as a *homeostatic* property cluster à la (Richard Boyd 1999; R. N. Boyd 1999). For example, Whitehouse-style social technologies would mean that the ritual infrastructure of religion directly upregulates cognitive religiosity and commitment to group beliefs. The mechanistic approach might therefore explain the bias toward grab-bags of religion-distinctive traits clustering together, either by joint causation or by those traits up-regulated one another. This 'causal network' view provides an alternative, more general gloss on the aspirational goal of (C. S. Alcorta and Sosis 2005) to identify the 'adaptive complex' that explains religion. The question then is how far out this causal network extends, and the degree to which religion-relevant evolutionary mechanisms are also cooperation-related (or are entangled with others that are more directly cooperation-related), both synchronically (speaking to stability) or historically (speaking to origin). Given the brief survey of mechanisms in the literature, religiosity and sociality might be very tightly integrated indeed.

From a philosophical perspective though there are two kinds of projects to be carried out: i) detailed and ii) big picture. The detailed projects are explorations of the explanatory power of

individual mechanisms and their likely effects (and side-effects) on social populations in the abstract. Big-picture projects may be compositional and syncretic: combining various plausible mechanisms into a causal/explanatory system, based both on what those mechanisms can do and on empirical constraints. Some approaches collapse the detail and big-picture projects – these are the single-mechanism explanations of religion which fit neatly into the categories I laid out in 1.2.1. The goal of this thesis is to make inroads into one 'detail' project: understanding the explanatory power of the central mechanism behind the signalling theory of religion. This is a mechanism which is sometimes viewed as a free-standing theory, but it is also capable of being combined with others (as Alcorta and Sosis suggest). Either way, 'detail' explanatory projects like this one must to speak to the big-picture constraints of empirical adequacy and fit with the facts as we know them – which in this case are the historical facts of human socio-religious evolution.

1.3. Timelines of emergence

We are now in a position to conclude the chapter by considering how explanatory approaches fare with regard to the available historical facts and the real target of inquiry: the evolutionary origins of the human cultural, socio-religious phenotype. In this last section I survey the state of knowledge about the origins of religion and complex human society and highlight the space for big-picture theorising relevant to the mechanism of religious signalling.

1.3.1. Uncertain origins

The question of how we got to where we are presupposes a point of departure: the 'there' from which we got to 'here'. The problem is that there are a number of 'there's we might start with. One might be the divergence of the *homo* and *pan* lineages. Another is the emergence of anatomically modern humans in the mid-to-late Pleistocene. However, these may have been complex, drawn out processes, for which the evidence is sparse, contested, and largely indirect. For example, the start of homo-pan divergence is sometimes placed at up to 13 million years ago, but with significant hybridization perhaps persisting until as recently as 4 million years (Patterson et al. 2006) according to 'molecular clock' evidence (though this is disputed (Wakeley 2008)). Anatomically modern humans (AMHs) were traditionally seen as emerging from more archaic human species about 200,000 years ago (200 kya) in East Africa with minimal morphological change since then, but a recent analysis of discoveries in Morocco suggests a more ancient and dispersed human ancestral population and a more gradual,

incremental evolutionary process (Hublin et al. 2017). AMHs were also once believed to have only left Africa and expanded into Eurasia (to mingle with Neanderthal, Denisovan, and surviving Erectine populations) about 60-80 kya, however earlier dispersals are now attested to[22] and the timing and mode(s) of dispersal look far more messy (Groucutt et al. 2015).

Regarding traces of religious activity, the evidence here is similar with respect to being both deeper in time and more uncertain than previously thought. As of a few decades ago, it was widely thought that religion was a relatively recent phenomenon, part of an apparently rapid cultural change during the Middle/Upper Palaeolithic transition (30-60 kya depending on geographical region) which Stephen Mithen in (Steven J. Mithen 1999) describes as a cultural 'big bang' and Jared Diamond called "the great leap forward" and a "magic moment in evolution" (J. Diamond 1991). Mithen argued that the entry into the archaeological record of sporadic ceremonial burial and other unusual material traces (such as the celebrated Hohlenstein-Stadel lion/man statuette) was evidence of a new religious phenomenon, rather than a lack of earlier preservation. In the last 20 years however, further discoveries have largely invalidated the 'big bang' idea, with more complete evidence smoothing out the rate of change and suggestive evidence of religion now going back much further and more broadly to include cousin hominin lineages. Finding include: symbolic use of ochres from least 100 kya (Henshilwood, d'Errico, and Watts 2009) with potential pigment use up to 1100 kya massively pre-dating AMHs (Beaumont and Bednarik 2013; I. Watts, Chazan, and Wilkins 2016), artefacts such as the Tan Tan figurine from at least 300 kya (Bednarik 2003), Neanderthal-attributed cave structures dated to 200 kya (Jaubert et al. 2016) and candidate symbolic traces (geometric scratch patterns) potentially dating back to about 500 kya and attributed to *Homo Erectus* (Joordens et al. 2015). With respect to burials, Mary Stiner recounts that burial indeed becomes very common after 28 kya but dozens of uncontroversial Middle Palaeolithic burials in Eurasia are attested (of both anatomically modern humans and Neanderthals) dating to 120 kya, with further candidate burials possibly as old as 250 kya (Stiner 2017). Many Neanderthal

[22] For example, AMH presence in Arabia and the Levant has been dated as far back as at least 177 kya (Hershkovitz et al. 2018), with dental and cranial remains suggesting such populations in East Asia by 100 kya (Liu et al. 2015; Li et al. 2017). However DNA evidence suggests a far more recent radiation event for the ancestors of modern non-African humans, who form a clade with Africans which excludes all ancient human specimens from outside Africa more than 50 kya years old (Skoglund and Mathieson 2018). This along with climactic evidence (Lamb et al. 2018) suggests that earlier migrating populations died out, though this too is disputed (Rabett 2018).

burials are very similar to later AMH burials (other than the bones themselves), with early AMH outside of Africa (120-90 kya) leaving a very similar material record with regard to technology (cultural/technological divergence only becomes apparent later). There is even evidence of multiple intentional body disposals by *Homo heidelbergensis* (350-450 kya) and *Homo nadeli* (~250 kya), though the intentions are of course debatable[23]. Further instances of apparent mortuary practices are detailed in (Pettitt 2015).

The *religious* interpretation of this data is even more contestable, both empirically and in principle. Ochre has mundane uses (Sterelny 2017b). Stiner is sceptical of inferring ceremonial burial from traces of flowers among burial remains. She also points out that extended periods of mourning and associated displays around deceased bodies have been observed among chimpanzees, bonobos, elephants, and dolphins (Elephants for example are known to revisit death sites many years later). Indeed, it would not be too great a stretch to see burying the body of a loved conspecific as simply a more ritualized extension of mourning, by a species with the technical capability for it (perhaps in order to simply protect the body from scavengers, as Stiner suggests). Sentimentality is not spirituality and symbolism is not metaphysics. Seeing such 'human' mental characteristics as a package deal is arguably what makes ideas like the cultural/cognitive 'big bang' more appealing.

With regards to religion then, it is difficult to know when something resembling it emerged. But over the last 200,000 years of the Pleistocene there was a slow increase in the archaeological signature of activities that we now take to be diagnostic of religious activities. That record becomes much clearer over the Holocene, when we also see clear, dramatic changes in how human societies were organised.

1.3.2. The Pleistocene-Holocene transition

The Holocene transition stands out in the archaeological record for several reasons. First, it was relatively recent and rapid, compared to the gradual change we see in the Pleistocene. Material traces of it are also obviously much better preserved, though the evidence we do have is geographically very uneven. It also transformed human society and lifeways in a number of

[23] There an illustrative debate regarding the deposition of the *Homo naledi* specimens in such an inaccessible location, and what this indicates with regard to their intent and awareness of death (Dirks et al. 2015; Val 2016; Durand 2017). One glib remark that can be made here is that serial killers, as with morticians, had to begin somewhere.

significant ways, reflected in the various aspect-specific terms applied to it: the Neolithic revolution, the agricultural revolution, the first demographic transition, and so-forth. It saw a dramatic change in technology, way of life, economic subsistence, with increased population densities and localisations of larger, sedentary groups with (eventually) hierarchical power structures. It also saw the rise early on of clearly identifiable traces of mass common religious activity. The main archeologically attested elements of the transition are summarised in table 1-4. For reasons to be discussed, each of these is surprising.

Table 1-4: Social, economic, and cultural changes characteristic of the Holocene transition

Before	*After*
Mobile bands of foragers	Sedentary farm/town societies
Small numbers of 'colleagues'	Large numbers (colleagues & strangers)
Opportunistic re-association/migration	Permanent association/low migration
Flat, egalitarian power structure	Structured hierarchies and networks
Policing of free riders and bullies	Elites escape egalitarian enforcement
Forager subsistence	Agricultural subsistence

For a sense of the timeframe, consider the span of recorded history, back to the first dynasty of ancient Egypt and the Sumerian city states in Mesopotamia. Less than halfway back to that 5 kya point is the career of Julius Caesar, with the Great Pyramid being more ancient to Caesar then Caesar is to us. With a similar ratio of antiquity (more ancient to the pyramid builders than the pyramid builders are to us) is the outbreak of the Holocene transition in the Fertile Crescent at about 11,600 years ago (the history of humans in Eurasia is at least five times more ancient). This is when the first towns and cities begin to appear in the archaeological record, as well as the advent of farming as the primary means of sustenance. Though we should be careful not to over-simplify, as the outward spread of Holocene societies from the Fertile crescent was slow and later centres of the transition around the world appeared to occur independently, it is true to a first approximation that these transitions were like a starter's gun that took us from Pleistocene society to human society as we see it now.

And it is difficult to overstate just how much human social worlds have changed. Insofar as we can reconstruct it, our Pleistocene ancestors lived in a social context which would seem very

strange to us now[24]. Up until the transition there was a great deal of social, cultural and technological development, but little obvious biological change or large-scale social revolutions. For the last 100kya at least we appear to have lived lives as mutualist foragers in mobile, ephemeral bands (or 'camps' (Marcus 2008)) of between 20 and 120 individuals; bands which were porous and regularly broke up and reformed with changing seasons, conditions, and relationships (Layton and O'Hara 2010). Taking contemporary forager societies as a model, it would have been possible to define larger cultural communities or meta-populations, via common language, ethnicity and practice, but there were no organised groups to speak of and no fixed abodes or territories to occupy: no chiefs, kings, tribes, or clans (at least not in any politically recognizable sense). We can think of meta-populations as being like vast decks of cards which periodically shuffled and dealt themselves into hands, the configurations of which culturally evolved in response to best ways of extracting resources from the environment, and how to those environments were treating them[25]. The larger aggregations were units of periodic genetic exchange (and between-band migration), information-sharing and dispute resolution, so they had some real social long-term significance, but they were not the immediate units of collective action.

This at least is the picture we get from the archaeological record and ethnographic studies of the few remaining forager societies (Bogucki 1999; Layton and O'Hara 2010), and it is in sharp contrast with that for other great ape species (where bands are more rigid and the community of potential band-mates is much smaller and more local). Such studies suggest that the band dynamic probably developed as a solution to the problem that hunting, a major source of food, was also a risky, scattershot endeavour: with low probability of high reward. Small hunting parties will usually return empty-handed, but when successful they have more meat than they could individually consume before it spoils. So, meat-sharing amongst a band of people (whom

[24] Though it is perhaps reasonably similar to some contemporary, ethnographically known societies. The degree to which surviving hunter-gatherer societies serve as a model for Pleistocene societies is an interesting issue. While basic uniformitarian assumptions appear valid, authors such as (Foley 1988) have noted that there are considerable dissimilarities in evolutionary ecology and corresponding strategies. The most obvious difference is that contemporary hunter gatherers mostly live off lands too marginal to have been appropriated for agriculture, or else are neighboured by farming societies – meaning they are not perfect models for the economic context of foragers in the highly fertile regions just prior to transition. These concerns I will set aside.

[25] Features like family groups and sex-based out-marriage norms of course mean that the shuffling wasn't entirely random, so this analogy should be taken with a grain of salt.

you can trust to reciprocate when they in turn are successful) allows for the optimum use of the fruits of your collective labour. Child-rearing follows a similar dynamic: risks are greatly reduced by relatively low-cost social insurance investments: roles and responsibilities are shared, reciprocated or socially pooled. Keeping the threat of free-riding to one side for the moment, the mutual benefit of "you scratch my back and I scratch yours" (especially if the transaction was more or less simultaneous) means this was (arguably) an evolutionarily expected strategy once we gained the emotional intelligence and communicative competence to negotiate it. And these fitness-enhancing skills and norms could flow more freely within a larger cultural population.

Another resource helping mandate this way of life was information, especially about food resources. Efficient foraging is about the marginal returns of gathering the lowest hanging fruit (as it were) and then breaking camp and moving on to the next area; perhaps between 5 and 80 times a year if current ethnography is anything to go by. The pooling of information about techniques for resource gathering is important, but so too is information about areas to exploit or move on towards. And this information pool is greatly improved by constantly refreshing the experience-base of the band. These are all reasons why the re-shuffling of band membership makes sense, and why there will be diminishing marginal returns (in terms of risk reduction and informational benefit) for continuing to increase the size of the band beyond a certain point.

A discussed by (Kelly 2013) with respect to contemporary forager populations, optimum interaction group size, relative mobility of those groups, and egalitarianism within them is greatly influenced by the distribution, variety and accessibility of resources. About 25 individuals per 'residential' band is common, with regional interacting populations (mediated by inter-band familial connections) of up to 800. The relative immediacy and labour-intensity of resource collection means that value and status to the group largely depends on what you do, not who you are or where you came from: much of the vaunted egalitarianism of hunter-gather society stems from meritocracy and self-reliance, with sharing of the surplus. Kelly also notes a tendency for egalitarian norms and conditional behaviour, for example to keep the egos of high performers (such as successful hunters) in check, but an inverse correlation with population and resource stress: in higher density populations and in harsh conditions the norms of society are less egalitarian. So, while the norms of such societies are by no means radically utopian (e.g. women had only incomplete autonomy), generosity is typically an expectation and the hierarchies and unequal relationships that do exist tend to be *personal* and lacking in

robustness both synchronically (no-one reliably dominates everyone) and diachronically (i.e. not reproduced inter-generationally).

All this is to summarise the state of human social lives by the time of the Holocene. Many interesting developments must have occurred to reach these already quite sophisticated levels of trust and cooperation, as cooperation among non-human great apes bears little resemblance (Tomasello and Vaish 2013). But given this picture of late Pleistocene life, the transition to Holocene societies like our own is far more of a challenge to explain. Whereas once we were mobile, we then settled down. Whereas once we lived in temporary groups where we all got to know each other, we then became permanently hemmed in by more-or-less trusted strangers (why?) in crowded settlements. Whereas once we would try our luck elsewhere if we were getting a raw deal, we then bought into the new settlement societies with all their risks and compliance costs. Whereas once we were egalitarian defenders of our fair share, the vast majority ended up toiling away as serfs (at best), redirecting a disproportionate share of production to their local lords. Whereas once we suppressed free riders, we somehow let them become the 1%. So much seems to have been given up in the transition that it looks like a strange career move for our ancestors to have made.

1.3.3. Possible explanations

If any period in human social history is crying out for an explanation then, it is the Holocene transition. On just about every contemporary view of it, the transition to farming involved cooperation and cohesion stresses – and an enhanced role for public religion is one plausible mechanism through which these stresses were relieved. I argue that there is a reasonably good case for this.

Prior to the work of a number of archaeologists in the 1960s[26], the dominant explanations were self-flatteringly progressive – the idea being that settled agricultural life was a straightforwardly rational improvement on forager life with more security, ease, and leisure (Barker 2006), and so simply inventing the technology was enough to kick off settled civilisation. However new ethnographic studies of hunter-gatherers like the !Kung-San put lie to this, as even in inhospitable environments foragers enjoyed relatively easy and secure lives compared to hard-toiling traditional farmers, with more healthy, diverse, and fungible food

[26] Famously culminating in the seminar and subsequent book “Man the Hunter” (Lee and DeVore 1968).

sources, and a lower parasite load. This comparison would have been much worse for early farmers with a much narrower range of and quality of domesticated crops, a fact attested to by skeletal and environmental evidence showing a marked decline in the quality of life for Holocene populations (Larsen 2006).

The gradual encroachment of farmers onto forager lands might be explicable militarily, via the order of magnitude greater increase in carrying capacity for land under the plough (J. M. Diamond 1998), a more rapid reproductive rate (due to giving away mobility), and the increased strategic value of land. Hundreds of sickly farmers defending fixed investments will still out-fight (in both strength and morale) dozens of relatively healthy foragers who have the option (for now) to hunt elsewhere. But the flip side of this is greater *vulnerability*: take away everything a forager owns and they are less well-equipped foragers, take everything from farmers and they cannot farm at all. At least at its inception then, obligatory farming seems an unlikely economic basis to organise society around.

And the idea that agriculture somehow opened the floodgates is further complicated by the fact that forms of agricultural technology and small, semi-permanent settlements predate the agricultural revolution proper by several thousand years. For example, the (currently) first attested bread-making is dated to 14,400ya, attributed to the Natufian culture in the pre-Holocene Levant, using a combination of foraged wild grains and root resources (Arranz-Otaegui et al. 2018), with evidence of cultivation ‘experiments’ at sites like Abu Hureyra at 12 kya or more (Moore, Hillman, and Legge 2000). Archaeologists appear to agree that limited cultivation and gathering of wild grains was used as a stopgap or back-up food supply for some foragers in the Fertile Crescent. Fixed settlements too had been used for thousands of years prior to them becoming permanent investments, as seasonal camps. Indeed, permanent habitation looks like it was a grim move that contributed to lowering health via accumulating waste, pestilence, and parasite load (Scott 2017). Again, technology does not look like the smoking gun of the Holocene transition.

This is not to say that agricultural investments, outputs and the changing economic and strategic environments they produced couldn’t have ratchetted up this aspect of the transition in some way. An explanatory narrative along these lines is certainly possible. But this bumps up against the basic demographic issue with reciprocity-based cooperation: bigger interaction groups are harder to stabilise. Large, dense populations of farming-supported settlers should have become

increasingly unruly and dangerous, especially if their wealth was in the form of hoardable, portable grain and had to be defended against opportunistic foragers and each other. The answer might conceivably lie with proto-state protection rackets started by rich elites surrounded by retainers or via patronage-client networks, but how did the elites escape social control in the first place? And, as (Scott 2017) and others point out, it actually took about 4,000 years of sedentism in the Fertile Crescent before the first definitive city states emerged. Meanwhile, the best explored archaeological site we have for an intervening settlement is Çatalhöyük (9,500 – 7,700 ya), which sported an urban population between 3,500 – 8,000 but with no evidence of elites (in the form of more ostentatious 'big man' dwellings or grave goods), giving "the overall impression is of a fierce egalitarianism" (Hodder 2014). In other words, many of the technologies significantly pre-date the demographic transition, but the demographic transition also seems to significantly pre-date the political transition.

One thing that we do see from the very inception of the transition though, and perhaps immediately prior to it (depending on the relative datings and interpretation of the evidence) is organised religion. Near the geographical centre of the Fertile Crescent's inferred origin, with inferred dates which tantalisingly match those expected for a 'smoking gun', like the Neolithic building complex at Göbekli Tepe.

This site is a series of temple-like structures ('enclosures') with distinctive T-shaped limestone monoliths, key-stone entry-ways, and decorative carvings featuring animal iconography (Schmidt 2000). Four enclosures have been excavated to date with organic samples (charcoal and organic fragments in preserved wall plaster) having calibrated radiocarbon dates between 11750 and 11310 years ago (Oliver Dietrich et al. 2013; O. Dietrich and Schmidt 2010). This is prior to generally accepted evidence of widespread cultivation or animal husbandry in the region (or anywhere else), and evidence of food remains at the site confirm a hunter-gather diet: with meat bones entirely coming from undomesticated game animals (mostly Persian gazelle and Aurochs) and only wild grains. There is no evidence so far of storage structures or domestic dwellings, or any other potential use of material utility (Peters et al. 2014).

These last points are worth emphasising: the best evidence is that the enclosures were not built by farmers, and they were built for a purpose that was not residential or of direct economic benefit. It is safe to assume a symbolic utility, but the symbolism is mysterious. The enclosures are often referred to as temples, but their interior space is too constrained to hold mass

ceremonies inside. They are not burial chambers either, but they were deliberately filled in and buried intact after a certain period of use – the in-fill including limestone fragments and broken bones, including some human (Schmidt 2010). Over about 1,300 years, successive structures were built over previous enclosures, over an area of perhaps 300 by 300 metres – there may as many as 20 enclosures in total to be excavated. This has been interpreted by the excavation team as a grand, symbolic 'burial and recreation' ritual repeated generationally (Oliver Dietrich, Heun, et al. 2012; Oliver Dietrich, Koksal-Schmidt, et al. 2012), and the enclosures interpreted as temples, with carved animals as guardians of some kind, and the T-shaped pillars as possibly anthropomorphic. As might be expected, the site has been the subject of much wider speculation[27], but a recent review of the more grounded literature can be found in (Henley 2018).

There are some fairly safe conclusions to draw, however. Regardless of what sort of rituals (if any) the temples/enclosures hosted, something culturally significant was happening at the site as their construction requiring coordination of labour and resources beyond those readily available to just a few loose bands of foragers or a handful of holy men[28]. The central monoliths of enclosure D, over 5m tall and weighing approximately 14.5t, were hewn out of a nearby limestone outcropping with granite tools and levered onsite by a large team of workers. Constructing entire enclosures (including detailed animal carvings) represented many months or years of full-time work, and probably generations of development with respect to engineering techniques. And if the purpose is religious it seems safe to date it further back than the enclosures themselves, as they are unlikely to be the first incarnation of whatever they were being used for. You don't start a religion by building its equivalent of the Sistine chapel.

---

[27] This includes the vulture iconography being related to afterlife and ritual sacrifice, symbolism of humans coming to dominate the natural world (Boric, 2014), as well as speculations about star alignments with Sirius (Magli, 2013), Deneb and the Cygnus rift (Lorenzis and Orofino, 2015). Speculative works like this are on a continuum which slides into outright crackpottery. It should be noted that all these speculations cherry-pick occasional symbolic motifs found at the site, none of which are present in all enclosures (except the T-shape of the pillars and the centrality of two pillars).

[28] Though see (Banning 2011) for a dissenting view: that a) the construction of at least the smallest enclosures was attainable by large forager bands, and b) that they served as extremely grand *houses* rather than public buildings (albeit with a striking density of symbolic content). As they could not house a whole band, they would presumably be 'VIP' residences, making this forager culture highly anomalous. Note also that this interpretation (unpopular as it is) is still not inconsistent with a religious justification for the VIPs, whose status (and symbolic trappings) would otherwise be mysterious.

The reasonable conclusion is that some sort of religion-like cultural phenomena was successfully and robustly demanding resource-intensive efforts for cultural infrastructure with no direct economic benefit, during the transition to settled society. The settling of the nearby agriculturally-supported proto-cities of Urfa and Nevalı Çori during the lifespan of Göbekli Tepe (with similar iconography and T-shaped pillars found) adds to the impression that something started there which radiated outward and helped seed the rest of the Holocene transition in this area. A few thousand years later we also see at Çatalhöyük rich and obvious remains of highly conserved and conformative religious and funerary practices, tightly integrated into the lives of the residents at the household level (S. J. Mithen 2004; Hodder 2014). So, while future discoveries might overturn this impression, *some* sort of 'religion did it' approach to cooperation in the Holocene transition seems plausible (at least in the Fertile Crescent), and reasonable motivation for exploring the explanatory possibilities in that space. At the very least, the Holocene transition seems to have often involved a much-expanded investment in religion-like activities at the same time as the social scale and temporal depth of collective action increased. Either way, the evidence is of a correlation between religion and cooperation during the transition, compatible with a broadly co-evolutionary relationship.

1.4. Summary and context

We can take a moment to review what these three strands of background literature collectively imply.

First, several straight-forward approaches to the cooperation problem don't seem to be a good fit for the human narrative (especially in the Holocene transition). For example, kin selection is does not fare well in this context, even with the aid of various social technologies. Keeping children close to the home community by giving them a clear lifepath into it would increase the chance that interactions between individuals are interactions between kin, but not in large communities, and traditional forager bands are not bands of close kin (Hill et al. 2011). Conditional altruism (mediated by norms) still requires reputation tracking, and the effective policing of shirkers and bullies places another, entirely non-environmental pressure on the upper limits of band size. To usefully monitor reputations via gossip (and the second-hand reliability of the gossip) is to track not just agents but their relationships (Dunbar, 1996), and this is a data set which increases exponentially (n(n-1)/2) with linear increase of group size. The hard limits of reputation-tracking in real-world settings are not straight forward, but

cognitive limitations would have been exceeded at some point beyond which an increasing proportion of agents are effectively strangers to each other. Anonymity and mysteriousness are the friends of the free-rider, so settled 'big societies' where free-riders can lurk in anonymity (and the reliability of gossip has lessened) cannot be straight-forwardly explained this way.

One big question here is the role and mode of inter-group dynamics, and the significance of group selection and cultural group selection. But forager interaction groups seem too porous and ephemeral relative to the time-horizons of biological group selection – out-marriage is also an important and functional source of migration (see also (T. Davis 2015)). Common cultural inheritance at a population level might instead allow us to see ethnolinguistic communities as units of selection within a meta-population. Any increased 'locking-in' of individuals within the community would also imply that such communities become less porous at their boundaries. The more that a community is kept in normative sync internally the more foreign, wayward or wicked that individuals from outside the community will appear. And the more elaborate and sophisticated a synchronising cultural edifice becomes, the more difficult and less profitable it becomes for immigrants or visitors to get up to speed. However, with cultural transmission in the mix there will be questions about whether evolution, growth, or cultural contagion are the better models by which to understand communities like this displacing one another. For example, the builders of Göbekli Tepe might have become dominant by virtue of being efficient users of human resources and out-competing their neighbours, or by others copying their strategies (or over-imitating their entire cultural package). If the later, was it because they were successful or because they were charismatic? In the absence of empirical evidence, there are still conceptual and modelling-based avenues of enquiry to pursue.

On intra-group level, more complex social technologies seem inevitable. This is where Whitehouse-style entrenchment of group norms might play a role, and/or signalling mechanisms and any other combination of cognitive or behavioural explanations. But these will need to all form part of a cultural 'scaffold' that is to some degree self-supporting with mutually reinforcing elements. Ritual practice centred around religion (or at least co-opted by it) looks like a reasonable bet. According to a scaffolding approach, complex socialised religion can be seen as largely concerned with community-building at multiple levels, whether by signalling or by normative reinforcement, tracking and shaping the social development of community members. Highly visible and endemic ritual practices, folk tales structured by a community-proprietary worldview, and regulation of sexual and parenting norms provide a

framework which usher children into the cultural/normative world of the community. Initiation and recognition rituals both provide direction at key stages of life and sort and subcategorise the cohort into distinct social roles and statuses with a sense of progression, destiny, and sunk cost. Some roles within the community might be singled out for special 'higher' statuses; but cemented as right and proper within the same progressive cultural edifice by a common narrative of being a people with a specific connection to places and resources, backed by cosmology and perhaps supernatural agency. At this point we have something that looks like religion, and a rough framework explanation.

Whatever the first-order content of the specific rituals and practices (and the more metaphysical elements of the grand narrative), in terms of their *function* these cultural primarily serve as technologies for the positive assortment of co-operators. Society-specific norms and expectations are indoctrinated and reinforced by pervasive ritualistic constructions and normative storytelling. Homogenisation of normative worldviews synchronise community members' expectations of one another, thereby reducing the scope for wasteful conflict. And mandatory participation in ritual periodically shines a public spotlight on individuals and their commitment to performing as expected.

That at least is an outline of the sort of picture that seems plausible, which combines biases, selection mechanisms, cultural transmission and so-forth. Religious life on this view was a process of cognitive and cultural niche construction.

What it takes to add precision to big-picture theory-building is more detailed understandings of the moving parts. Significant work has been done for example on the psychological effects of ritual and religion with respect to identity fusion and group ideological cohesion. Likewise, cultural transmission mechanisms such as CREDs are being studied as paradigms for religious learning, taboo practices and in behaviour change. To some degree, religious signalling theory is being tested in observational studies, but (as I shall argue) there is considerable work to be done with regards to the mechanisms of signalling and the explanatory work they can actually do.

2. Big Gods

In the previous chapter I made the contrast between explanatory mechanisms, and 'big picture' theories which combine and contextualise mechanisms into overall explanations of phenomena. In this chapter, I look at one of these big picture views that has recently risen to prominence: the so-called 'Big Gods' theory. Though it harkens back to earlier explanatory approaches, the modern Big Gods theory has been developed over the last decade by Ara Norenzayan (a social psychologist of religion) and co-authors, especially in the book "Big Gods: How Religion Transformed Cooperation and Conflict" (Norenzayan 2013), and more recently in a multi-author target article in Behavioural and Brain Sciences (Norenzayan et al. 2016a). The book and its theory received wide attention, being the subject of several symposia including via the International Cognition and Culture Institute, and the journals Religion, and Religion, Brain & Behaviour. These, and the commentaries attached to the target article, as well as other literature (including a feature in Science magazine (Wade 2015)), make this view one of the most well chewed-over and influential in recent years. While the theory does not make explicit use of the signalling mechanism, and has some weaknesses that I will be elaborating upon, it is also my intent to use it as a template for theories of an appropriate scope, complexity, and ambition to explain socio-religious changes such as we see in the Pleistocene-Holocene transition.

2.1. Overview of the big gods model: theory and mechanisms

Despite the diversity of response and rapidly developing literature, the Big Gods theory is relatively simple and intuitive plausible. The conceptual starting point is the similar to the 'fear of supernatural punishment' mechanism (FSP) postulated and developed in (Johnson and Krüger 2004) and (Johnson and Bering 2006), and discussed in the previous chapter. I.e. fear of supernatural punishment generates a kind of religiously conditioned self-restraint; where an indoctrinated belief in supernatural punishment outsources the policing of behaviour to the believer. You believe the gods are watching you, so you think twice about cheating or stealing (i.e. defecting), given the expected consequences of their righteous anger. Where Norenzayan and co-authors differ from the earlier uses of FSP is i) by looking to more group-level and fitness-orthogonal evolutionary mechanisms to upregulate the beliefs generating the FSP effect, and ii) gathering a large body of empirical evidence to back up the relevant mechanistic claims.

The theory involves a story about how such 'big gods' beliefs might co-evolve in a mutually reinforcing relationship with a broader, more generalised pro-sociality, and with social groupings that are larger, more complex and more cohesive or culturally centralised. In brief, the behavioural restraint that FSP generates expands the circle of trust and cooperation (since the god-fearing agents are more trustworthy and cooperative). This allows larger, more culturally coordinated societies, which in turn are more effective in transmitting and reinforcing the FSP-friendly beliefs – thereby forming a virtuous cycle (or vicious, depending on your point of view). Mechanisms such as CREDs and cultural group selection (and less enthusiastically, signalling) are then appealed to as ways to drive this process along.

Figure 2-1: Schematic representation of posited causal processes in the Big Gods theory

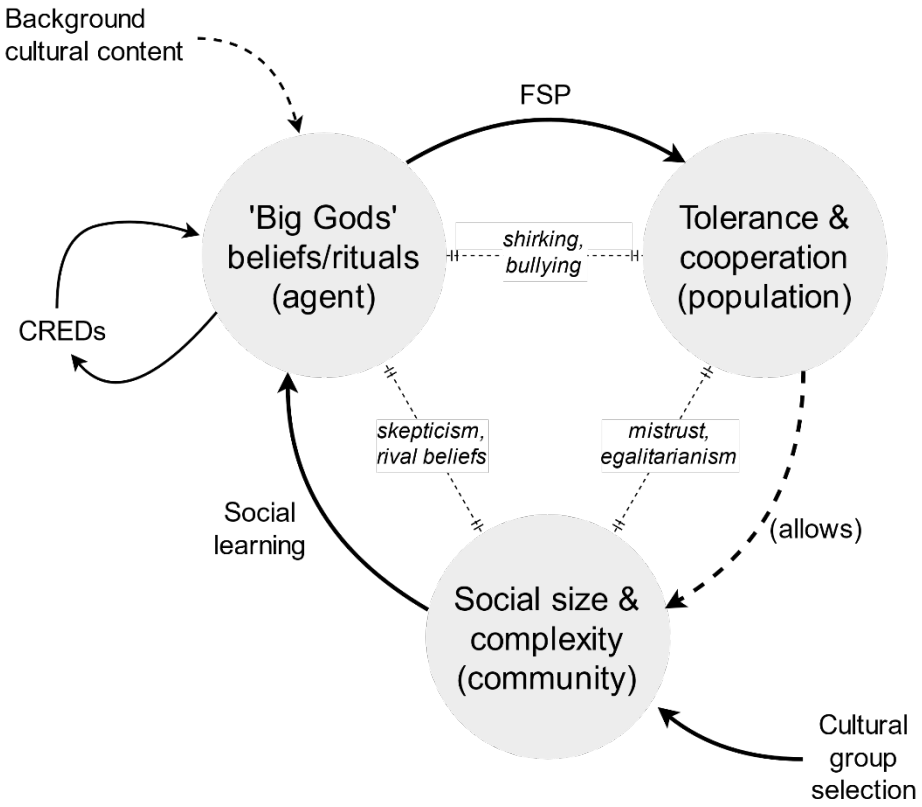


Figure 2-1 is a proposed schematisation of the theory that I will be taking as an illustrative reconstruction of how it works[29]. Starting at the top left of the diagram: background/ancestral cultural content may include beliefs and personal rituals directed toward supernatural forces or

[29] Note that this is my reconstruction based on (Norenzayan et al. 2016a). Proponents of the view might well quibble with this, but the intent here is to make a plausible case for the view, and to highlight the moving parts of it that are the most interesting for current purposes.

entities which (to some degree) oversee human affairs and mete out rewards and punishments. These are 'Big Gods' beliefs. Rational agents who have such beliefs should both exhibit more prosocial self-policing (through fear of supernatural punishment, FSP), and have more confidence that the actions of others are being policed, which should together 'upregulate' the average level of cooperation and tolerance of strangers and elites within the population. The strength of this 'FSP' causal upregulation arrow would depend on the prevalence and motivating power of the beliefs. This works maladaptively (with respect to the individual) against pragmatic incentives to shirk, exploit, bully, or withdraw cooperation, but it *allows* larger and more complex (and inegalitarian) communities to be stable. Such 'big societies', are then also apt for upregulation by cultural group selection, in the case that competitive advantages such as economic specialisation, economies of scale, and military capability are realised. The final link in the cycle is that successful societies are better at transmitting their cultural contents to agents (future generations, migrants, members of rival communities, etc) via social learning. This upregulates Big Gods beliefs and rituals, which would otherwise be opposed by general scepticism, resource expenditure, and competition with the beliefs of rival communities. This is further helped along by agent-to-agent, fitness-orthogonal transmission mechanisms such as CREDs (see 1.2.3) – which will be all the more credible from members of a successful community[30].

This causal process would be an evolutionary engine that ratchets up the strength of the cycle's agent-population-community nodes, pulling the society in question clear from the pressures and incentives which enforced the old Pleistocene equilibrium. It gradually generates stronger and stronger beliefs in more and more demanding deities, and bigger, more intricate and more demanding cultural structures which increasingly squeeze and overpower the prior social predilections of the people who became caught up in it – all the while vacuuming up more and more people, territory, and resources. In this way, social religions and social complexity are modelled as co-evolving through mutual reinforcement, with socialised religion playing a necessary role in the emergence of Holocene-style societies, but without unexplained prior (question-begging) existence.

---

[30] Especially, one might add, if those members have taken advantage of the new scope for inegalitarianism by appropriating the social trappings of the religious life; such as a priestly class or elites who sponsor and/or are endorsed by them.

This is therefore an elegant and promising model, which offers a plausible and robust co-evolutionary mechanism that goes well beyond ‘because religion’ pseudo-explanations of complex, Holocene-style societies. How this might be dovetailed with the archaeological evidence to make a convincing narrative is another question. Perhaps the Big Gods mechanism was already at work in the Pleistocene but was held back by the limited resource base or population density of the forager economy (or until the end of the Younger Dryas removed the brakes?). Perhaps it works best as a relatively novel and rapid ‘outbreak’; a stochastic event the result of a mutation in religious culture at an opportune time, when the big gods causal complex could end up co-opting ‘emergency use’ agricultural technology to feed an increasing hunger for labour, food and other resources. As the model is essentially a feedback loop, it is a tantalising possibility for explaining the demographic transition as a runaway process. Regardless, a co-evolutionary model of this scope and ambition seems necessary to do justice to any narrative with a significant ‘religion-did-it’ component.

2.2. Predictions and evidence

In his book Norenzayan argues that Göbekli Tepe provides indicative evidence of a Big Gods religion in effect during the Holocene transition, siding with the interpretation that at some of the symbolic content depicts abstract anthropomorphic god-figures (albeit cryptically). Notably, in their target article, Norenzayan and co-authors also appeal to (Bellah 2011)’s analysis of religious beliefs up to and including the empires of the Axial age, citing his correlation of social size and characterisation of ‘intermediate’ chiefdom-based societies: “[t]heir gods are more powerful and moralizing than those of foragers, but not as full-fledged as the Big Gods of states and empires” (Norenzayan et al. 2016a, 9–10). Notably though, other analyses argue that “most gods of early large-scale societies were not concerned whether one behaved in prosocial ways” (Baumard and Boyer 2014), and the target article was accused of ignoring this (Boyer and Baumard 2016). Regardless of interpretation, this illustrates a general prediction that will need to bear up if the theory is correct: for the Big Gods evolutionary engine to be explanatory it will need to be the case that ‘big societies’ actually did tend to have religions with powerful moralising gods (especially while those societies were ‘ratchetting up’, so to speak). Of course, in the absence of actual texts, archaeological signatures of religious content are particularly difficult to infer. But other historical and contemporary (traditional) cultures should also evidence these correlations.

That then is one basis on which to test the Big Gods theory. However, the fit-to-world of a big picture composite model is only one condition for it to be satisfactory. The more basic questions are in the details, i.e. do the component mechanisms function the way they are supposed to, and what does they evidence say about them in isolation? Because some of the big-picture predictions are influenced by the more detailed level, I will consider first the evidence concerning the key psychological mechanisms of the Big Gods theory, and then return to the big picture.

2.2.1. Evidence for the watcher effect mechanism

The Big Gods theory relies on claims regarding the social psychology of religion and cooperation which make some obvious predictions, and Norenzayan's book is largely concerned with findings in social psychology which shore those up (including his own work and that of his colleagues). While I called first link in the Big Gods model a 'fear of supernatural punishment' mechanism, Norenzayan's originally characterised the psychological basis of this in the "watcher effect": that watched people are better behaved (prosocially speaking). The effect extends to inanimate cues with no obvious watcher involved: for example the famous Bateson honesty box experiment (Bateson, Nettle, and Roberts 2006). This uncontrolled experiment was carried out in a university faculty tearoom with an honesty box for staff to pay into when adding milk to their tea and coffee: photocopied pictures of eyes placed above a honesty box were superior to photocopied pictures of flowers by a factor of three in terms of how much money was coughed up. Similar results are found in controlled lab experiments involving public goods games played over a computer terminal: participants are more generous or fair in their offers when there's a visible avatar or even inert stylised eye glyphs drawn in crude dots (Haley and Fessler 2005; Rigdon et al. 2009)[31]. The hypothesis that connects this to the Big Gods theory is that belief in gods (that might be watching us) triggers the watcher effect and generates similar pro-social behaviour.

We should therefore be able to see this effect when comparing the behaviour of believers and non-believers. In this area though the evidence is more equivocal. For example, according to Norenzayan's own review of the prior evidence, no strong correlation between generosity and

[31] It should also be noted that there are doubts about the replicability of such findings (Sparks and Barclay 2015; Matsugasaki, Tsukamoto, and Ohtsubo 2015; Dear, Dutton, and Fox 2019).

merely *holding* religious beliefs shows up in standard University-based studies of public goods games[32]. This adds to a mixed body of evidence (summaries of this literature include (Batson, Schoenrade, and Ventis 1993; Malhotra 2010)), which casts doubt on the long history of assumptions in this regard and much evidence of *self-reported* correlation[33]. More supportive (though still equivocal) is the literature on the role of religious *cues* in generating the so-called "Sunday effect" (Malhotra 2010; Norenzayan 2014; Xygalatas 2013). The Sunday effect as reported by (Malhotra 2010) is that religious people in the USA are significantly more generous in making charitable donations but *only* soon after attending church services. Malhotra's interpretation is religious beliefs themselves are relatively inert without activating cues, and governing norms much be made salient. This is also argued in (Shariff and Norenzayan 2007), where participants in a public goods game were 'primed' ahead of participation using scripts with implicit religious cues, resulting in elevated prosocial performance. A later meta-analysis of similar studies agreed that these findings were robust (Shariff et al. 2016). However, a larger registered replication study of (Shariff and Norenzayan 2007) with updated experimental and analytic methods and n = 650 participants (approximately ten times more than in the first paper) found no significant watcher effect (Gomes and McCullough 2015). The authors of the meta-analysis have expressed scepticism with that result and suggested that the lack of findings was due to differences in experimental design and/or statistical methods (Willard, Shariff, and Norenzayan 2016), however for the moment the general findings for this and other priming effects are uncertain (Van Elk et al. 2015; Rivers and Sherman 2018).

There are two areas of evidence which do support a connection, however. First, some indirect support is found for increasing levels of religiosity being predictive of in-group cooperation, lower rates of free-riding, or cultural stability in studies comparing religious to non-religious

---

[32] such as Dictator or Ultimatum games, where one participant of a pair gets to divide up of a sum of money between the two, and their partner either has no say or gets to veto the distribution entirely.

[33] One famous early study is the 'Good Samaritan' study of (Darley and Batson 1973). This study set up seminary students to walk past an actor professing distress while on their way to an engagement, with a key manipulation being whether or not that engagement was to deliver a short talk on the parable of the Good Samaritan. Neither this manipulation nor religious personality measures predicted the likelihood of the student stopping to offer assistance, though there was some correlation with between religious measures and the nature of help when it was offered: subjects identified as exhibiting strong doctrinal orthodoxy tended to be more persistent in offering help despite refusal (dubbed 'superhelpers') and less sensitive to the actor/victim's own characterisation of his needs. In line with similar findings, this did not support religious priming as a behavioural driver, but suggested religiosity might influence specific behaviours, given the "time and scope to permit personality characteristics to shape them" (Darley and Batson 1973, 108).

communities such as 19th century American communes (Sosis and Alcorta 2003; Sosis and Bressler 2003), contemporary Israeli Kibbutz (Sosis and Ruffle 2003) and Indian Madrassahs (Ahmed 2009). More directly, adherence to Abrahamic religions in non-western or traditional societies *has* been found to predict a higher rate of altruism in ultimatum/dictator games, adding 6-10c per $1 to split offers (Henrich et al. 2005). While small, this effect is significantly higher than western-based studies. Norenzayan's position is that this body of evidence shows that people from Western, Educated, Industrialized, Rich, and Democratic (WEIRD) societies, who are the usual subjects of university-based psychological studies, have another cultural layer (secular civil society) which masks the association between religiosity and cooperation and shouldn't be taken as representative of ancestral minds (Henrich, Heine, and Norenzayan 2010). A more recent synthesis of cross-cultural studies using public goods games suggested a more nuanced and split effect: with big god beliefs helping in *extending* cooperation beyond immediate kith and kin, and "supernatural punishment clearly had a more consistent, robust, and significant effect than monitoring" (Purzycki et al. 2018).

2.2.2. Considering the details

This rapid overview of the evidence is of course insufficient to make any grand assessment of the watcher/FSP effect. For example, it may be that the effects are real, but the effect sizes are problematically small in populations that are easy to study. Evolution can still do a lot with a small effect – just add time. But these considerations also identify a conceptual concern: there is a liberal amount of interpretation required in order to get experiments to say what the BG theory needs them to say.

For example, in the case of the watcher effect, the exact mechanism of action for the eye cues is not clear. The idea that they trigger a deep "I'm being watched" response, and fear of punishment is an interpretive hypothesis. Perhaps gaze detection makes us feel watched and making us feel watched makes us cautious – there is a reasonable fear among social animals of stepping out of line, and the eye cues might indeed hack into an adaptation of that evolutionary origin. *Or*, perhaps what's being cued is a sense of merely being in the *presence* of someone else: being in a social situation rather than a "lone scavenger"-type situation. On this interpretation it might not be a fear of punishment but rather some sort of coalitional instinct that's being hacked; which is the sort of thing that might already be reasonable to assume from evolved Pleistocene-type cooperation. The evidence is compatible with both interpretations, but only the "I'm being watched" one dovetails with the Big Gods theory. The fact that I just

completely speculatively invented a vaguely plausible alternative mechanism, which may or may not be realistic, is in part beside the point and in part illustrates the point – the more interpretation you have on the path between data and conclusion, the less robust the conclusion.

There is also a fair amount of interpretation required in the case of the religious priming data. Even if these results turn out to be valid, it is not clear if it is Big-Gods religiosity being triggered or some other psychological mechanism. For the WEIRD vs non-WEIRD comparison, it would have to be the case that i) there is good explanation for civil secular society's masking effect (and to Norenzayan's credit he makes a good case for this), but at the same time ii) there are no rival explanations for the (still small) priming-generosity correlations in the non-WEIRD setting. As another example of a common-cause alternative interpretation: perhaps in more traditional societies with a deeper integration of religion and society, the people who are less religious are less religious because they just don't like people as much. Antisociality (or social alienation) would then be the common cause of both lower religiosity and lower cooperation, which wouldn't support the causal link required[34].

None of this is to say that the psychological evidence goes *against* the BG theory or that it is been misused in any way; just that it is not clear how to assess the degree of support conveyed, and it is reasonable to retain a degree of scepticism. Big Gods might be consistent with the evidence, but that evidence is far from being conclusive in its favour. And, though the empirical argy-bargy here might be more complicated than I have done justice to, there is at least a mild philosophical point that can be made: there is a lot of interpretation standing between the data and the theory.

2.2.3. Cultural comparisons

The Big Gods model relies on other causal-explanatory mechanisms, such as cultural group selection and CREDs, which I briefly discuss elsewhere and will not explore further here. Instead I will now turn to the more significant empirical prediction: that there should be a correlation between the presence (or history) of moralising high gods in a society, and markers of social cohesion, scale, or internal complexity. Some sort of pretext for this has already been mentioned via the commune studies (Sosis and Bressler 2003; Sosis and Ruffle 2003), however

[34] It is annoying for experimentalists to have armchair commentators pull out random what-ifs like this, but if the results are being used to push a much more wide-reaching and abstract conclusion (and the evidence for the more proximate mechanism is cloudy), then a bit of counter-speculation isn't entirely out of place.

for Norenzayan's purpose these are only suggestive (as it is difficult to control for other variables in communes founded within specific political and cultural settings). In his book, he further cites cross-cultural correlational studies based on ethnographic databases of societies around the world, coded for belief in 'moralising high gods' (MHG) and measures of social complexity (i.e. testing the correlation between big gods and big societies).

However the methodology of these studies have been criticised, for example by (Atkinson, Latham, and Watts 2014) who write "[a] major problem with almost all of the cross-cultural analyses of MHG data is that the statistical tests employed assume that the data points are independent when they are not." For example, there needs to be careful handling of societies which have been influenced by European or Arab colonialization, which introduced Abrahamic cultural/religious lineages on one hand and (in some cases) also introduced structured and centralised social institutions. To include such societies would be in effect to massively 'double-count' the Abrahamic lineage, but once this is controlled for (according to (Atkinson, Latham, and Watts 2014)) the supporting correlations drop below significance: only within the Abrahamic tradition is there positive evidence of a correlation.

If this result holds, then it would be problematic for several reasons which have not been fully articulated previously. First, the Abrahamic tradition is only a few thousand years old, so its success in no way supports a *deep* prehistorical role for big gods in shaping Holocene societies from the very beginning. Second, as a single data point it might just as well support any other common feature of the Abrahamic religions as the causal factor, not just belief in the Abrahamic god. A strong, state-linked priestly class, the concept of 'the elect', reliance on set scriptures, prophet figures (and so forth) might equally well explain the success of these religions, as far as we know. Yahweh/God/Allah might be the hero of the story, but he might not be the reason why the story is so successful.

The success of the Abrahamic religions might also have nothing to do with the religions themselves, and more to do with their being part of the longest incrementally accumulated body of culture in the world. Farming, the first cities, states and empires originated in the ancient, well-developed, and culturally rich Middle East; and the polities which later promulgated Christianity and Islam are the inheritors of that. There is a parallel here with the way that various Mesopotamian conventions such as the Zodiac, and the 360-degree circle piggy-backed their way into wider use via a) the ebb and flow of empires which co-opted them and b) sheer

temporal priority. So too, perhaps it should be expected that *some* religion from this region would be the 'break out' religion. It might be a geopolitical accident that *this* was the religious tradition of the region which happened to be carried out to the wider world, by descendent cultures with an economic and technological 'head start' over others, or who just happened to be in the right place at the right time. The English language for example has had a globally successful radiation comparable to that of the Abrahamic religions, but no one would seriously look to its intrinsic linguistic features for an explanation of that success. Accident should be the null hypothesis here, and a sceptical eye should be cast on any claim that the successful *earned* their success by virtue of any inherent quality (especially if it's a quality which the claimant particularly favours).

So, while they provide a template for the Big Gods idea (and are of great contemporary political relevance) the strong signal from Abrahamic religions might be leading us astray. Status as global religions is not evidential, and we still need either i) a better statistical test of the global success of Big Gods religions, or ii) an independent reason to think that the Big Gods engine is effective. Regarding the second point, it is possible to debate the role of the Big Gods cultural trait in the incubation and growth of the Abrahamic religions *within* Middle Eastern-descendant cultures (including European and Arab cultures). In (Norenzayan et al. 2016a) for example the authors cite literature on the spread of Christianity in the early Roman Empire. This is a possible strategy, but the story often told about this has little to do with theological aspects of the Christian 'cult'. In an oft-cited account, (Stark 1996) for example points to the specific norms that were enforced, not the motivations of those enforcing them[35]. Demonstrating the opposite would again require evidence of the watcher/FSP effect. In absence of the proximate mechanism evidence, supporting a particular story would require a different dataset which spoke to *specific* historical processes insular to the cultures of origin. There are simply too many alternative historical hypotheses for the grass-roots radiation of Christianity to supply favourable evidence for the Big Gods theory. Even then, this would not licence extrapolation back to the Holocene transition or more generally.

[35] For Stark the important normative differences were the Christian ban on an otherwise-common Roman practice of female infanticide (giving Christians an advantage in purely demographic terms, and in conversion via exogamous marriage), and principles of mutual charity in need (serving as a survival-enhancing social insurance policy).

We would either need to see the Big Gods effect in action in situ, multiple times, or a correlational analysis of outcomes which demonstrates its effect among a sample of cultures where radiation and cross-contamination is carefully controlled for. Thus far we have neither. There is also at least some evidence that the opposite is true, though (again) it is disputed. In (J. Watts, Greenhill, et al. 2015) the authors use a linguistically inferred cultural phylogeny of Austronesian cultures to test for correlations between the presence of 'moralising high gods' and the presence of political hierarchies (as a proxy for social complexity). Their findings were that the correlation was weak, and statistical analysis with respect to likely ancestral priority (which the phylogenetic analysis allowed) suggested that political hierarchies tended to precede moralising high gods, rather than the other way around. This was followed by another study which indicated that human sacrifice was a far better predictor (than moralising high gods) of the evolution of social complexity (J. Watts et al. 2016).

While these studies are statistically sophisticated, there is some wiggle-room for the Big Gods theory. First, they are all from within a single cultural lineage, meaning there may be common 'defeater' effects for the Big Gods model which are not reproduced more broadly. Second, test variables were drawn from available codings in the Polutu database (J. Watts, Sheehan, et al. 2015), with 'moralising high gods' used as representative of Big God deities, where 'high god' includes being a creator god. Big Gods proponents might object that this is not close enough a match (Big Gods need not be creators). Finally, what these studies *do* show to be predictive of social complexity is what the authors call "broad supernatural punishment": the belief in any supernatural agent or process that "reliably monitors and punishes selfish actions". This will be discussed further the next section, but the later target article gestures at a more liberal notion of the watcher effect and/or supernatural punishment which incorporates Karma (to some extent), and also potentially mana and tapu (Norenzayan et al. 2016a, 8), concepts of moral consequence from many of the cultures that Polutu covers. This permits a response along the lines that the Watts et al. result might instead suggest some degree of confirmation for the general framework, with mana and tapu acting as impersonal surrogates for Big Gods.

Overall, the evidential support for the Big Gods theory is suggestive but ambiguous (at least to an outside observer). However much of this ambiguity stems from exactly what is being claimed, how vital it is to the theory, and how it should all be tested. Which vehicles of supernatural punishment or moral consequence does the Big Gods theory include, other than just the Abrahamic prototype? This is an area where greater clarity is required.

2.3. Connecting the conceptual and the empirical

While it is implied in Norenzayan's original exposition, the target article makes it explicit that big god beliefs admit of degrees. They posit a *continuum* of big god belief types which contribute to expansion of cooperative groups via self-monitoring, but in proportion to how closely they approach the Big Gods paradigm. So, there is now a continuum or spectrum of 'Big-Goddishness', with the canonical big gods such as the Abrahamic god at the most extreme end, and along which other beliefs can be arrayed. The approach here makes sense as it seems to address both theoretical and empirical needs; however I argue that it raises further issues.

2.3.1. Operationalising the Big Gods continuum

First, a spectrum or continuum in no way an ad hoc move. Big God beliefs are evolved traits of a culture; therefore, there must be an evolutionary pathway to such beliefs from the 'no gods at all' trait. They need to be able to have emerged more or less incrementally out of other traits that are only incompletely 'Big-Goddish' or not Big-Goddish at all; this is implicit in the 'ratchetting up' metaphor and therefore just part of the view. Indeed, in this regard the authors are on a firm footing: we know that the Big Gods cultural trait has arisen in a demonstrably incremental manner in at least one case (as the authors describe; the early Abrahamic god Yahweh began as just one tribal god of the Israelites, among several).

It is illustrative here to distinguish two broad processes for modelling the increase of a trait within a population, which I gestured at earlier when outlining my scepticism about the Abrahamic data point. One process would be to imagine agents with traits coded as continuous variables, with the variables of succeeding generations of agents[36] depending (in part) on the success of the previous generation. If the trait conveys a fitness advantage in a given environment, then even if it begins set at near zero, we can expect to see a gradual increase in the relevant variables over the whole population over time (as long as there is a mechanism to introduce 'noise' or some other source of variation for it to become present at low levels). Yahweh's evolution from god to God is the paradigm example of this. The alternative is to imagine discrete traits carried by alleles: a fully formed trait that is advantageous might spread

[36] The generations of an agent-based model formal modelling can refer to reproductive/biological generations of agents with 'baked in' traits that do not change during their lifetime, or alternatively to a population 'decision state' with new 'generations' occurring when individual agents change their trait. I use the term here purely in a formal manner, i.e. ambiguously between biological generation and decision state.

to fixation by enhancing the relative fitness of its carriers (even in low initial numbers), allowing them to invade the population and out-compete their rivals. This might be one-way belief in a mature Abrahamic God might spread within a previous isolated population, from an initial 'infection' event. These are two different mechanisms of spread; the first should be kept in mind when the Yahweh example is appealed to – where the belief trait is endemic but is evolves along a temporal gradient toward a more 'full-blooded' deity. Other historical examples cited by the authors, such as the spread of Christianity within the Roman Empire, are more apt for the invasion dynamic: a fully-fledged trait introduced into a population and proliferating to fixation as its carriers reproduce or convert non-believers.

Importantly, both processes model selection at the individual level for explaining intra-cultural changes and are distinct from group selection mechanisms operating at an inter-cultural or meta-population level. In the Abrahamic narrative, group selection processes might (perhaps) be apt for trying to explain the *survival* of the Abrahamic religion's early incarnation via Israelite culture (as opposed to its Phoenician, Egyptian or Mesopotamian regional rivals), and/or the spread of a later one via the vehicle of the first Caliphate. While the paper is necessarily abstract and the authors do not explicitly individuate these mechanisms, a systematic application of the view to the history of human religions would have to be careful about the evolutionary processes attributed to each component of the narrative.

And emphasising a continuum with a temporal gradient opens up an interesting response to the critical ethnographic data discussed earlier, i.e. (Atkinson, Latham, and Watts 2014) and (J. Watts, Greenhill, et al. 2015). These studies use binary coding of cultural traits – for example there being 'Moralising High Gods' or not – with the cultures coded positively occurred in relatively small numbers (the trait was surprisingly rare). Norenzayan et al do not say as much, but it is open to them to dispute binary coding as distortive and insufficiently subtle: contested examples and 'not-so-big gods', such as the gods of pagan Rome, still count as 'big-goddish'-*enough* for their purposes and should ideally be included in a model that 'scores' beliefs in a continuous manner (as well as including greater or lesser degrees of social cohesion). While this may not convince their critics, it at least points out a degree of mismatch between theory and criticism. The problem then would be one of empirical tractability – it is not obvious that the ethnographic data is rich and detailed enough.

Because we know that there must be a spectrum, we know that the various cultural incarnations of Mars and Demeter must fall along it somewhere. The only questions are where they fall, and whether time gradients along this spectrum correlate with changing social cohesion in a way that confirms or disconfirms the Big Gods hypothesis: do societies get more cohesive and complex as their gods get bigger? The most appropriate test of the hypothesis would be snapshot studies of how cultures change over time, but again evidence we have simply isn't fine-grained enough for this.

2.3.2. Issues with the continuum

So, we have both an evidence gap and questions about empirical tractability. The beauty of straight-forwardly scoring a culture as 'big gods present' or not is that it is indeed straight forward in a way that incremental scoring isn't. But there is also a conceptual problem with any continuum proposal: what would it be for a culture to be exactly 80% along the big god spectrum? If the authors' intentions were to rise above certain criticisms, this has not yet been achieved. The low-resolution tests that went against the Big Gods theory will still stand unless the higher-resolution ones which the authors gesture at can be operationalised. Individual historical correlations are of course available to be appealed to – but given the richness of human cultural history this will always be open to charges of cherry-picking. Genuinely robust empirical evidence of causation (as opposed to psychological plausibility) is still lacking and (if anything) we have gone backward because we don't know how to operationalise the variable of interest. Beyond the basic notion of a continuum, we have nothing specific to guide us in this respect.

For useful operationalization we would need a better understanding of the theoretical basis of the continuum, its dimensions and their relationship to each other, and ultimately an index apt for proportionality comparisons. If we considering moralising, punishing gods in the way the theory characterises them, then there should be at least three components to the continuum, one for each of the key 'bigness' properties of big gods: i) their degree of omniscience in monitoring our actions, ii) their degree of commitment toward moralising/punishing, and iii) their power to punish (i.e. how bad a punishment is expected). These are three orthogonal traits which (conceptually speaking) can vary independently depending on the personality of the projected deity. So understood, any empirically tractable operationalization of the big gods effect would have to decompose it along these three dimensions, perhaps with weightings or combination/multiplier effects, and further weightings for the cultural embeddedness or

eminence of the belief in the mind of the believer etc. This would not be impossible, and as a way of conceptualising the continuum it is reminiscent of the 'hypercube' dimensional analyses of (Mitchell 2000). But empirically speaking it would will be time-consuming and fiddly, requiring case-by-case interpretation on ethnographic databases where much of the relevant data might never have been collected, and it would be vulnerable at every turn to charges of subjectivity. Especially given the level of dispute over the relatively simple coding schemes already in use, this does not look like a promising research project.

And translating such a 3-dimensional score into behaviour would not be straight-forward. How would intermediate deities register an effect, and how to calibrate this? For example, a hand-wringing moralising god who sees everything but has with no power to punish might engender some sort of respect from believers, but if fear of supernatural punishment is doing the psychological work then these deities probably won't have much effect on the behaviour of ne'er-do-wells. Once every god has been scored from zero to one on the three scales of interest, it might be tempting to just multiply those numbers to get their 'Bigness' fraction. But consider three gods: one all-knowing and all-powerful but only partially interested in your doing the right thing, another who is entirely interested and all-knowing but only partially able to punish, and a third whose imperfection is that they only know what you're doing some of the time. Reasonable believers might fear these intermediate gods in quite different ways, with different attendant behavioural effects. It would be useful to know much more about the psychology of god-fear.

All this is not so much a serious proposal as just an illustration of the explanatory gaps in the Big Gods theory. The notion of supernatural monitoring and punishment is intuitive but underspecified, and this only becomes clear if the tri-omni Abrahamic God is taken off the table as a starting point. The point is that you can't simply 'scale down' the all-knowing, all-powerful, all-judging Abrahamic god to use as a conceptual element of a theory; not if *fractional* Big-Goddishness is supposed to be empirically tractable. It would be deceptive to ignore the can of worms that this opens. Much more work would need to be done, and it's not clear (to me at least) what that work would look like. And until such clarity in operationalisation has been achieved, the model has decidedly *not* positioned itself to address the empirical concerns discussed earlier.

In laying out this worry I of course oversimplified, by talking about divine beings as the sole source of supernatural punishment. This is not the case in target article presentation of the theory. The supernatural punishment/reward systems of non-Big God religions (e.g. Hinduism, Buddhism, Jainism) are indeed treated as part of the general framework and "reflect historical convergences between religion and public morality, although the precise psychological mechanisms are not as well understood as for the Abrahamic religions" (Norenzayan et al. 2016a, 9). It is the authors intention (thought they admit that there is much work to be done to develop this properly) that beliefs such as karma would also be part of the Big Gods effect in some sense; likewise, Oceanian concepts of mana and tapu. Conceptually speaking, this backtracks somewhat from big gods and the watcher effect to a broader conception of supernatural punishment: to 'cosmic justice' in general, with big gods as just particularly agent-based manifestations of this. As already mentioned, while the authors have not stated as much, this move may also provide some basis for pushing back against the findings in (J. Watts, Greenhill, et al. 2015), as this study was conducted in an Oceanian setting where moralising high gods play second fiddle to more abstract mechanisms of mana and tapu, arguably vectors supernatural punishment, moral consequence, or cosmic justice. The predictor in this setting should now be some disjunction of moralising high gods and other cosmic-justice beliefs, so a failed test of one disjunct alone might not be cause for alarm.

This is another alteration which is sensible in principle but pushes the theory further from clarity with respect to empirical tractability. Buddhism for example is probably the only other good case of a 'break-out' religion: originating in India but colonising much of East Asia and spreading as far as Japan. It also underpins social structures where it dominates yet is mostly non-theistic – so it stands out as a challenge to the Big Gods theory if not incorporated somehow. But if operationalisation of orthogonal theistic properties was difficult, operationalisation of a disjunction of theistic properties and non-theistic cosmic-justice concepts seems even more ambitious. Should we posit a fourth dimension to the continuum, from the impersonal (Karma) to personal (grumpy old Yahweh)? If so, where do multiple gods fit on that scale? What about the more esoteric God of contemporary monotheism, or of Spinoza? Does this dimension matter much at all (the success of Buddhism might suggest not) and how does it interact with the others? The "how big a god" interpretive issues are also reintroduced in this context, with the authors arguing that many apparently non-Big Gods religions only have this characteristic at the refined, theological level, and for the everyday

lives of believers “anthropomorphic beings often reappear with a vengeance” (Norenzayan et al. 2016b, 49).

Again, this is not a straw-man demand for a fully specific model. Models always over-simplify, so while (for example) just treating size of god as single measurable dimension might do much violence to the difference in theology, it might work fine as an operationalising and modelling strategy. Coding ethnographic data as well is always a matter of interpretation. But if we are going to code and simplify some variables for statistical testing, we need some sort of broad agreement on what sort of methods would be acceptable. What is the appropriate kind of ethnographic dataset? Could you use a disjunctive binary variable e.g. big god, or karma, or tapu, etc.? A weighted variable (or variables)? If so on what principle? Or perhaps a proxy of some kind: e.g. whether is there the local equivalent of the sentiment “what goes around comes around”. What we need is some better sense of what sort of footprint or signature we should expect to see in the cross-cultural data if the theory were true. Until then, dialog on this issue is all too likely to resemble researchers like Watts et al being accused of attacking straw men, while Norenzayan et al are accused of shifting goal posts[37].

2.4. Conclusions

The Big Gods theory seemingly makes predictions which are only ambiguously supported by the psychological literature (with respect to the main mechanism), and only ambiguously supported by cross-cultural studies (with respect to the big-picture predictions). Both are further complicated by a difficult problem of operationalisation. Given that the big gods continuum and its putative psychological mechanism are the defining features of the Big Gods theory, I think the question marks over them are problematic (at least for now), and reason enough to pursue other options in the meantime.

Again, the criticisms here are not decisive or fatal. And it is important to properly frame the philosophical criticism of the Big Gods continuum: there is nothing wrong qua model with invoking what turns out to be a structured, multi-dimensional continuum. There is good theoretical basis for this in (Mitchell 2000), and Peter Godfrey Smith for example uses a similar continuum framework with respect to the structured sub-components of natural selection

[37] This summarises an exchange that the author witnessed at workshop in May 2018.

(heredity, sensitivity to fitness landscape, etc) where they are visualised as orthogonal dimensions which define a possibility space for populations to be more or less paradigmatically Darwinian (Godfrey-Smith 2009). However the analogy is an imperfect one. First, the Big Gods continuum has nothing like the cache or pre-established plausibility base of natural selection: added complexity and structure in its case will incur far more of a cost in terms of overall parsimony. Second, while the Godfrey-Smith project is largely a philosophical and conceptual exercise, the Big Gods project is explicitly an explanation to be applied to and empirically tested by specific archaeological and anthropological evidence. It therefore needs to be robustly operationalised and made empirically tractable if it is to achieve its intended purpose, and it has not been.

I think it is therefore fair to ask: is the fear of supernatural punishment worth all this effort, and what would happen if we were to discard it? I think we can still retain something like the basic structure of the model (as I sketched it out). Recall that the Holocene transition problem is a problem with the formal modelling of cooperation. This is the body of theory for which the transition is an anomaly. The Big Gods model would have explained a great deal and has the right level of ambition, but it is a psychologically specific model that carries a lot of baggage with respect to interpretive assumptions and empirical tractability.

Simply on grounds of parsimony, what would be preferable is a religion-cooperation co-evolutionary approach which makes fewer assumptions, requires less interpretation, and instead invokes more abstract, ‘multiply realisable’ mechanisms with independent justification. Religious signalling is one such possibility, and in the rest of this dissertation I propose to see how far this approach can be taken.

3. Rationale for a signalling theory of religion

In this chapter I turn to the signalling theory (or, rather, theories) of religion. Thus far I have laid out the case for a co-evolutionary explanation of religion and cooperation (especially in the Holocene transition) and looked at the Big Gods theory as one example of that. The overall narrative here is one of incrementally building a case. While the Big Gods theory meets many of the desiderata for a co-evolutionary explanation, it is not ideally *efficient* in this regard – its postulated causal mechanisms appeal to features of human psychology of questionable robustness and universality, and which themselves require some sort of explanation. The overall empirical case is a long way short of being decisive. One way of 'doing better' than Big Gods (all else being equal) would be to do away with the need for such explanatory dependencies and rely instead upon causal mechanisms with fewer commitments regarding human cognitive idiosyncrasies. As I will argue, signalling theories of religion (at least some of them) are able to make good on this strategy, and do useful explanatory work at a level of description that has less exposure to the empirical fortunes of cognitive science. The goals of this and the following chapters is to flesh out and justify that claim, and to flesh out the explanatory scope of different models of signalling in religion.

In this chapter, I defend a *theory schema* for signalling theories of religion, based on a putative explanatory functional role that religious signalling had in the evolution of both religion and cooperation. Chapter 4 looks at how signalling works, and how signalling theory (especially from biology and formal modelling) refines both the scope and explanatory potential of a signalling theory of religion. The thesis then further drills down into the specific context of religious signalling builds the case for realistic applications of the theory to religious ritual and other situations (chapter 5), with subsequent chapters devoted to testing the developed claims using formal/computational methods and discussing the results and overall prospects for the approach.

So (at last), what is religious signalling theory (RST)? The basic hypothesis can be stated relatively simply: participating in religion signals the cooperative 'social quality' of the participant to conspecifics, and this drives the evolution of both the signal and social quality/cooperation. This theory has it that rituals or other visible religious tropes of the right sort will allow the community to expose uncooperative and under-committed individuals, allowing for those individuals to be excluded from cooperative ventures and the public goods

which flow from them. On some versions of it (costly signalling), the ‘right sort’ of rituals are those which impose costs upon participants, but which fall more harshly upon the uncooperative and under-committed, such that they are disproportionately ‘priced out’ of participation. Assuming adaptation has occurred, the participants that remain would more likely to be genuinely safe to cooperate with, and so the prospects for cooperation-sustaining positive assortment of ‘team players’ are improved. In other versions, costs are not necessary (or at least not central) and difficulty gradients or constraints on signalling do the filtering job instead. Different versions might coexist and might also complement more cognitive models (e.g. if rituals both impose costs with the right profile and trigger mechanisms of social bonding and prosociality), but the general function of signalling is to improve prosocial positive assortment. In any case, in at least some versions, RST offers the tantalising promise of a plausible, self-contained co-evolutionary explanation which does not rely on leveraging conveniently pre-evolved psychological traits, and which is largely tractable using formal methods.

This is an initial, abstract characterisation of the general approach, but the position of RST in the literature is complicated. The central idea is intuitive, but there are devils in the details. What constitutes a signal? What are the benefits of signals to senders and responders? Must they 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’ have not always been 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 the broad predictions of RST (Power 2017; Sosis 2003; Sosis and 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 (Henrich 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 and Sosis 2011). As introduced in chapter one, CREDs are costly displays of sincerity 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, the displays are clearly signals in some sense, but is it of the same kind as other RST theories? There is of course a semantic or ‘conceptual housekeeping’ issue here. But beyond this, the issue is

whether RST is best seen as a disjunctive collection of proposals, or whether different combinations of signalling features (e.g. cost, interpretation, honesty, benefit, mechanism of action) constitute importantly distinct explanatory kinds. Is there scientifically substantive dispute going on here (because CREDs and other signals work in a fundamentally different way) or is it just terminological?

I argue that there is a unified explanatory notion of signalling to be pinned down (cohering with my initial abstract characterisation), which justifies substantive scientific distinction-making. This is not a novel view, but the main contribution of these central chapters will be a stepwise abstract reappraisal of this notion from 'within' signalling theory. This is intended as a contrast and compliment to more traditional surveys of the literature, e.g. (J. H. Shaver and Bulbulia 2016), rather than a repeat of such work, and comparative empirical adequacy will likewise not be discussed in any depth. The first goal is to sketch out a 'core' schematic for understanding RST, outlining its distinctive commitments and explanatory virtues. This will (I hope) be ecumenical and recognisable enough to include most putative signalling theories of religion, while also serving as a basis for demarcating signalling theories proper from rivals. The following chapters then delve deeper into the available options for RSTs: different models of signalling, and templates for bringing this theoretical machinery to bear on putative real-world religious signalling phenomena.

3.1. The Signalling theory as an explanation

Often going under the name of the 'costly' or 'honest' signalling theory of religion, the contemporary notion of RST as a causal explanation is most closely associated with the work of Joseph Bulbulia, Richard Sosis, and co-authors (Bulbulia 2004a; J. H. Shaver and Bulbulia 2016; Bulbulia and Sosis 2011; Sosis 2003; Sosis and Alcorta 2003). Earlier articulations include those by anthropologist William Irons (Irons 1996; 2001) and others such as (Cronk 1994), again in the context of more general proposals about the biological evolution of moral traits and institutions with prior champions (e.g. (Alexander 1987)).

However, this literature has had other sources of influence. One clear precedent is the theory of signalling in animal communication, including signals as manipulation (Richard Dawkins and Krebs 1978; Krebs and Dawkins 1984), and the so-called handicap principle of costly signalling (Grafen 1990a; Zahavi 1975; Zahavi and Zahavi 1997), with signalling in biology (especially ecology) emerging as a complex field of study into its own right (Jay M. Biernaskie,

Perry, and Grafen 2018; Hebets et al. 2016; Hebets and Papaj 2005; Hurd and Enquist 2005). A relevant parallel literature on costly signalling also exists in economics, beginning with Michael Spence's Nobel prize-winning model of job-market signalling (Spence 1973), with economic analyses of the organisational benefits of costly religious demands considered in economic analyses. For example, in (Iannaccone 1992) entry costs for communities are given an economic analysis, and this is a commonly-cited paper in the religious signalling literature. Another oft-cited influence that I will return to is Robert Frank's theory of emotional displays as evolved signals to 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 signalling supposedly works, and over what scope and timescale. For example, Sosis sees religious signalling as pro-social mechanism for large groups (with personal reputations being more important in small groups), while for Bulbulia signalling can only act in a one-on-one manner on a local, observable scale. And while the biological and formal signalling models which act as a precedent are also typically based on *population thinking* (e.g. modelling signalling via dyadic interactions within populations of more-or-less interchangeable individuals), more complex verbal models of social/network configuration have also been proposed e.g. top-down, one-to-many religious leadership or 'charismatic' signalling (Bulbulia 2010; John H. Shaver, Fraser, and Bulbulia 2016), and ritualised, aposematic signalling of power and brutality (either top-down or between-groups) (Bulbulia et al. 2017). Such different approaches, along with differing appeals to specific psychological/cognitive traits, 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.

The literature on signalling theory in religion is therefore far from monolithic. Consequently, it would be difficult to draw a systematic, scholarly characterisation of RST which captures each and every 'signalling theory of religion'. There are, in effect, a variety of different RST-like theories, with different core assumptions, commitments, and explanatory targets. Different ideas about the relative importance of these (with respect to what it takes for a theory of religion to be a *signalling* theory) will specify distinct (though overlapping) sets of theories for

inclusion into the RST concept. Any attempt to characterise a core notion of the RST will therefore be revisionary to some degree, and my own proposal here is no exception.

I base my characterisation of the RST on the current literature on signalling and signalling theory with respect to evolutionary biology, philosophy of biology, and formal modelling. To a first approximation, signalling theory in the style of Irons, Sosis & Bulbulia can be seen as a conceptual 'porting-across' of animal signalling models to human religious expression, and this is indeed how it is often described, especially with respect to costly signalling (John H. Shaver, Fraser, and Bulbulia 2016). The approach here will therefore be abstract and analytic rather than an exercise in curve-fitting or characterisation. The intent is not to tailor an umbrella concept for all those theories which have been labelled with the term 'signalling', but rather to isolate the distinctively evolutionary strain of signalling explanation that many of these views recognisably appeal to.

In this chapter then I focus on a central theme: a simple evolutionary picture as the *characteristic core* of the RST approach, with more developed theories seen as variations on the theme. This move would also help to isolate why we might be interested in a signalling theory of religion in the first place. As suggested in the introduction, signalling theory supplies a way in which religion and sociality might have co-evolved in an explanatory package deal; analogous to co-evolved trait-pairings in biology such as the striking morphology of flowering plants on one hand and the foraging behaviour of pollinating insects on the other. The 'core' theme of the RST also has several features which make it particularly interesting in the abstract, and in the remainder of this section I will develop these ideas in parallel with elaborating the proposed core theme of RST.

One specific dimension of difference that I will bracket off until the next chapter though is the more 'formal' matter of how signals are kept honest. Many discussions about signalling theory in religion begin by appealing to costly signalling, the handicap principle, and/or index signals, and offer examples of seemingly irrational religious costs or emotional displays that appear to fit formal signalling models used elsewhere (especially in biology). I will be arguing that the formal picture is much more complicated than this, and in any case secondary to the business of characterising RST itself.

3.2. Mechanism, endogeny, and scalability

We can better understand core RST in comparison with the Big Gods theory from chapter 2 (Norenzayan 2013; Norenzayan et al. 2016a). Though Big Gods was developed as a sophisticated combination of causal mechanisms (including signalling-like mechanisms), the characteristic component is that beliefs in potent, morally attuned supernatural entities (or other forces of cosmic justice) serve to upregulate pro-social behaviour among believers. Call this the 'fear of supernatural punishment' (FSP) mechanism. In contrast, RST is a partner choice mechanism: the religious signal improves the positive assortment of compatibly prosocial agents[38]. If religious prosociality somehow significantly increased the likelihood of being surrounded by likeminded types (instead of those who would merely exploit you), then it *would* pay. On this view, religious participation in one's community evolved as a commitment signal for community-inclined individuals to collectively condition their behaviour upon, allowing them to preferentially congregate for mutual benefit and more reliably identify and exclude exploitative free-riders and unprofitable cooperation partners. Puzzling features of religion (notably its demandingness) might become explicable as adaptive traits to better facilitate the commitment signal and/or ensure its reliability (though see chapter 4).

An important explanatory advantage shared by both views is that they go beyond the usual 'dyadic' model of individual interactions, and out-source the maintenance of cooperative behaviour. As discussed in 1.1.3, the famous, purely dyadic model for stabilizing cooperation is reciprocal altruism, which can serve as something of a baseline for other explanations to supplement or improve on. One pragmatic limitation on reciprocal altruism was the need to gather and utilize a quantity of information, which scales up dramatically as more potential interaction partners are considered. In any case the theoretical limitation is clear: for any given background rate of actual trustworthiness and risks/costs of exploitation, there will be

[38] To a first approximation, this 'compatibly prosocial' can just mean just a simple altruistic trait. Altruists do better when positively assorted (as long as the overall benefits of cooperation outweigh the costs), as they all share in the benefits. But we can obviously think of more detailed forms of behavioural compatibility. Consider for example be the ability and willingness to cooperate on a limited range of projects and/or according to specific norms and procedures, such that only specific behavioural trait-profiles will generate net benefits when brought into association. The positive assortment and signalling rationale can therefore apply more broadly than to just simple, idealised altruists; facilitating the assortment of compatible and complimentary behavioural phenotypes in general.

informational requirements that limit reciprocal altruism's ability to stabilise social worlds, beyond a certain size.

The Big Gods FSP mechanism would impact on reciprocal altruism by improving the background rate, so mitigate the group-size limitation. Believing that bad behaviour will be punished makes group members less likely to defect, so big gods should tend to promote bigger societies. The RST mechanism instead supplements interaction histories with a more directly accessible body of information – visible conformity with religious norms which (as honest signals of prosocial commitment) correlate with a lower likelihood of defection. Indeed, if this signalling is i) reliably correlated with pro-sociality and ii) highly visible, then the need to monitor potential interaction partners for pro-sociality might disappear entirely, and cooperative societies facilitated by religious signalling could grow arbitrarily large. So, FSP would raise the size of groups that our cognitive-informational capacities would allow reciprocal altruism to stabilise; with indefinite group size possible if absolutely *everyone* has been scared straight. In contrast, RST is an alternative conditionalization strategy which in idealised form (with honest signals visible on demand) would make reciprocal altruism redundant.

It is worth noting that the high visibility condition is not universally accepted as being met. Bulbulia for example sees the observability of religious commitment signals as problematically limited, and therefore favours a small-scale role for religious signalling (Bulbulia 2010). Bulbulia has in mind discrete ritual displays, and there are indeed limitations here. Rituals are public, communal participation displays, so more observable than cooperation history, but simply observing and tracking prior ritual participation would indeed impose similar cognitive demands. But for our purposes there is also room to consider *persistent* commitment signals, such as conformity to restrictive codes of dress, language, behaviour, or association. These would be observable and understandable by cultural conspecifics even in the absence of any personal familiarity. Ostensibly cheap secondary signals – e.g. scarification or wearable status markers earned after passing initiation rituals – might also reliably and persistently advertise prior commitment demonstrations. This is especially plausible if religious communities have packages of prescribed acts and permitted signifiers that are policed by the community, 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. impersonating a high-status figure within a well-policed community). These ideas will be returned to in chapter 6, but for the moment (despite

Bulbulia's reasonable concerns) I will make two in-principle assumptions: a) RST can stabilise cooperation on a larger scale than reciprocal altruism, and b) an extended sense of religious signalling might be visible enough to help stabilise cooperation even in large social worlds.

In summary, while RST and Big Gods both propose a co-evolutionary explanation of religion and cooperation, the nature of the explanation differs. Big Gods sees big society-cultivating religion evolving via cognitive biases and transmission mechanisms that are themselves up-regulated in big societies, but which are fitness-orthogonal (and likely maladaptive). I.e. religion evolved in spite of its individual-level fitness costs. The RST on the other hand has religion's evolution being driven by individual-level selection and *because* of its costs, insofar as those costs underwrite the honesty of coordination-facilitating signals and help solve the cooperation problem[39].

These features of core RST lend it some degree of attractiveness, at least in the abstract. Specifically:

1. It would operate at the same level of explanation as the problem it would address (i.e. individual-level fitness).
2. It offers to by-pass the need for specific, personalised knowledge about potential cooperation partners, permitting arbitrary large societies to evolve.
3. It has relatively few explanatory moving parts in relation to its explanatory payoff. It doesn't need to appeal to any exogenous, pre-existing cognitive biases, belief structures, or maladaptive traits (though these may be included in more detailed variations of it).

That third (supposed) theoretical virtue deserves some unpacking. There is of course nothing scientifically dubious about appealing to deeply entrenched, empirically sound human traits, nor should such appeals exclude a theory from the RST tent. Indeed, some influential signalling theorists stress the importance of specific belief structures (Bulbulia 2010; John H. Shaver, Fraser, and Bulbulia 2016), or see signalling as likely just one component amongst many within rich, heterogeneous causal complexes that explain religion (C. S. Alcorta and Sosis 2005). Instead, the motivations here concern parsimony, and categorisation.

---

[39] Because they work in such different ways, it is important to note that these two effects are in no way mutually exclusive – both might operate at the same time.

The parsimony point rests on idea that, all else being equal, simpler and more self-contained explanations should be preferred, as every additional appeal to cognitive biases or other background traits raises further questions. Most obviously, empirically well-established cognitive features will have their own origin stories to tell, and the more that they are appealed to in an explanation of a socio-religious phenotype the more that their own evolution and stability become the real, ultimate explanation of the initial explanatory target. This is especially pressing if religion is seen as maladaptive: why was the direction of evolution toward the new (socio-religious) traits and not away from the old? As observed by (Bowles and Gintis 2011) in the context of the evolution of cooperation, the reality of a behavioural bias (social preference) which *would* be explanatory if it had been continuously present over the relevant evolutionary timescale is not by itself explanatory: you also need to explain where it came from and why it was stable over that timescale.

For example, I argued in chapter 2 that the ubiquity of the cognitive profile required by the big gods model is not uncontroversial. It is not clear how typical it is of current human phenotypes, and even if it were, it is unclear how deep in history such a profile might have formed. To some extent then, the Big Gods explanation of religion is hostage to the timelines and explanations of the cognitive features of humans that explain it – it raises questions as well as (potentially) answering them. Of course, nothing here suggests that answers to these questions would not possible, just that they *will* be required, and an idealised RST theory does not share such requirements to that degree.

3.3. Functionalism and the scope for explanation

What I have done so far is sketch out an abstract 'theory schema' for RST, and this characterises RST as a broadly 'functionalist' explanation. But a functionalist explanation of what? With respect to the definitional matters discussed in the first chapter (Harrison 2006), the general explanatory target of the schema, the 'religion' that the RST would explain, is the social phenomenon of religion. That is to say: why beliefs and practices exist as socialised, locally common features at the community level, in virtue of their community-binding function. It therefore sits most naturally within a broadly Durkheimian tradition focused on religious practices as central to group cohesion (Durkheim 1912; Watson-Jones and Legare 2016; Whitehouse and Lanman 2014). Religious signals facilitate and assist in maintaining community cohesion.

3.3.1. Functionalism, flexibility and diversity

One in-principle implication here is that RST might be able to remain largely neutral on the content of the beliefs and practices in question. It is the signalling *role* of religion which makes it stable and ubiquitous, not any specific religious form or content (though this is often re-inserted by RST theorists e.g. via the influence of specific belief types and 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 moralising 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 signalling (generating a coordination signal for prosocial types) being multiply realisable.

There is an obvious analogy here with language. Given its broad utility in solving coordination problems, it is not entirely surprising that language is ubiquitous among language-capable human beings, nor that linguistic populations are typically much larger than local interaction groups. Debates about universal linguistic features aside (see for example (Berwick and Chomsky 2016; Planer 2017)), there is nothing much beyond the mechanics of sound production that maintains linguistic commonality between, distant, causally unrelated populations – hence the drift and diversity of human languages across the globe. Another obvious analogy is currency. A currency (in the minimal sense) is anything that functions as a currency: something that has a significant enough local tradition of being adopted and honoured as a medium of exchange. So, it can both be true that currency is ubiquitous but also that local currencies differ widely from place to place, variously based on precious metals, institutional guarantees, seashells, cigarettes, or blockchain. Language and currency are both multiply realisable functional devices for the solution of coordination problems; so too (says core RST) for religion.

Both advantages and disadvantages accompany this high degree of flexibility and diversity. Like languages, RST-explicable religions could take on an almost arbitrary variety of forms and content (in principle). With this comes potential applicability (of more tightly specified RST variants) to a variety of explanatory contexts, including ritual vs other religious tropes, small-scale vs large-scale societies, and different explanatory timescales from deep-time cognitive & cultural evolution, to social transitions in the Pleistocene and Holocene, to specific application to historical eras (such as the axial age) or contemporary religious settings.

However, insofar as human religions do come with highly typical contents or cognitive traits (e.g. beliefs in supernatural agency), RST alone might have few resources to explain them.

Importantly though, RST is potentially quite modular, and its multiple-realizability is non-exclusive. Nothing prohibits distinct cultural/ritual features from serving parallel signalling functions, or single cultural/religious tropes and practices from having distinct signalling functions in concurrent effect. And nothing prohibits any of these from combining and interacting with other causal mechanisms (for example with more explicit content biases). So, while the T in RST stands for 'theory', the real explanatory object being assessed is the RST mechanism, or rather mechanisms. It is a feature of functionalist theories in general that they permit explanatory pluralism and we should therefore be cautious about talking about 'the' signalling theory or signalling mechanism. The RST theory schema will of course require fleshing in with more formally specified signalling models (chapters 4 and 5) and some plausible assignments of those models to real-world religious target systems (chapter 6). But the point for now is that, as an explanatory approach, RST need not specify (and be slave to) a single applied narrative or overall explanatory package.

3.3.2. The costs of religion

When characterising RST, I said that it can render religious costs explicable as signal costs, which can help guarantee signal honesty (though this is a complicated issue that will be an ongoing theme in future chapters). Indeed, 'costly signalling' versions of RST have been explicitly proposed for helping explain the costliness of so much religious ritual and other demands on religious adherents (C. Alcorta 2017; Bulbulia 2004b; 2004a; Sosis 2003; Sosis and Alcorta 2003). On this view, costs are not so much a bug of religion as a feature.

And there is ample documentation of such 'irrationally' demanding practices. Much ritual is physiologically demanding, involving uncomfortable fixed positions, or ritual movement that is not trivial in either energy expenditure or investment in proper training and practice (Bulbulia 2004a). The most extreme practices such as cutting and piercing are typically only occasionally demanded, but even innocuous ritual observance generally involves tithing significant time and resources which could otherwise be spent productively. An example of the former kind is the Shia Tatbir or Talwar Zani; a bloody and dramatic ritual in which Shia men use swords to cut their own heads in a symbolic, violent spilling of blood, but which is controversial among religious authorities, practiced by only small minority Shia Muslims, and then only at a specific

annual event. At the other end of the scale is the familiar pan-Abrahamic tradition of sequestering a certain amount time every week on a holy day for religious observance. While it is rare for ritual demands to be *cripplingly* onerous, they are typically non-trivial especially in their traditional context: removing a full day from one's economic week for example was a far more significant a cost/risk in pre-modern or traditional societies than it may appear from a WEIRD (western, educated, industrialised, rich & democratic) perspective. And, as seen in the case of Göbekli Tepe, the costs of religious monuments in pre-modern society must have often been a very significant fraction of the social surplus. For costly signalling theories of religion (as the name suggests) these seemingly irrational costs have a job to do: they permit the coalescence of a cooperative society and therefore the emergence of its benefits.

Again, other explanations of religion and social hierarchies (e.g. as outlined in chapter 1) can make sense of a certain amount of policing and punishment. Highly intense, aversive ritual has powerful motivational effects which might help explain (at least in a proximate sense) the voluntary submission to such rituals without any direct benefit to either participant or community. The idea of human beings as *homo economicus* should also not be overplayed – maladaptive explanations of religious practices are always live options. So, it is not as though signalling theory as outlined here is the only game in town, and it would be a mistake to simply cite the existence of costly religious practices as evidence of costly signalling (another point that will be returned to in more detail later). But the sheer diversity of costly demands we seen in religious practice around the world is again something that teases for a broader explanation. A functionalist role for costs *in general* is therefore worth looking into, as this again implies a multiple-realizability which might be filled by a great diversity of costly activities. Costly variants of signalling theory might fit that bill[40].

3.4. Summary

The proposal therefore is that 'core' RST mechanisms offer individual-level, fitness-driven signalling explanations of the evolution of the socio-religious phenotype. The more *endogenous* this process is theorised to be (i.e. the less that it leans on other causally relevant features of ancestral humans or their environments), the closer it is to core RST. Variations on the RST theme might have it that signalling was not the only driver, or that specific cognitive

[40] As well might e.g. Henrich's CREDs theory, which obviously also has cost explanation baked into it.

features of human beings interacted with signalling to accelerate or drive it toward specific manifestations. But the more necessary or integral those signalling-exogenous features are to the causal mechanisms posited, the more licenced we are to say that such a theory is only marginally a signalling theory or not an example of one at all. On this characterisation, the previously mentioned CREDs theory would fall outside of RST proper, as CREDs do not improve the fitness of those who generate them. However, the 'charismatic signalling' of Bulbulia and co-authors seems like a variation on the theme. It is in this sense that the core RST framework offers a way to help categorise and clarify the literature in a principled manner – that is, once greater clarity is achieved on available signalling models and what an evolutionary account of them requires.

However, any classificatory utility here (which would be contestable based on a more in-depth survey of the literature) is a side-benefit, and the main point of this chapter is a more modest one. All I have argued is that there exists a clear explanatory approach, arguably common to much of the RST literature, which, when isolated, appears to make promising predictions with regard to how human societies were shaped and scaled up, and displays certain abstract virtues with respect to explanatory parsimony. These virtues in and of themselves are of course no evidence of actual explanatory utility or empirical adequacy, and certainly not explanatory superiority. Indeed, they might instead help explain the appeal of signalling theory to some theorists in the *absence* of empirical adequacy and explanatory superiority. Clarifying the idea in the broadest possible terms has been the only goal.

With this schema in mind, different possible versions of RST can be seen permutations of the different ways of filling in the details: who the agents are, which formal models of signalling are matched to the relevant realiser mechanisms (e.g. actual displays, their costs, and the emotions, or cognitive biases which drive them), and (though this is less commonly considered) how the standard moving parts of a fitness-driven evolutionary explanation (units of selection, heredity, variation, and differential fitness) are specified. Because at this level of detail the pluripotency of core RST 'in the abstract' becomes subject to constraints and begins to make more specific predictions.

My goals in the following chapters are therefore more involved but still relatively modest. I will be elaborating on what an evolutionary account of signalling involves. I will argue that the menu of available signalling models is diverse but disjunctive: each model comes with its own

quirks and apt domain of applicability. No single model delivers all the explanatory benefits we might be interested in, but in concert they may recover much of the potential outlined in this chapter. I aim however to be more specific: how might we best apply these models, and in what sort of scenarios and social institutions might signalling models be recognisably active? What would we predict their effects to be? I will also be bringing formal methods and computational models to bear on that last question. This will provide more robust proof of concept (and in some cases refutation) of some of the ways in which relevant signalling models (with idealisations relaxed in sensible, case-specific ways, and operating within plausible parameter ranges) might help explain the evolution of religion and human society.

4. Signalling theory for religious signalling

Chapter three introduced a theory schema for the signalling theory of religion, which included the proviso that religious signals be "reliably correlated with pro-sociality". There are many ways to ensure reliable correlation, and in this chapter these will be explored. That is to say, what are signals and how are they kept honest? Without honesty (as opposed to sincerity) the characteristic fitness advantage of RST disappears. Altruists benefit from assorting with other altruists, not just those who believe they are altruists and make a song and dance about it. In the previous chapter I deliberately avoided the usual tropes at this level of detail (such as 'costly signalling'), since there are several importantly distinct honest signalling models to consider, all of which are in some sense secondary to the core idea of RST as just outlined. Crucially, different models of signalling will meet the theory schema's requirements and desiderata to different degrees – and none of them simultaneously. More comprehensive surveys of the modelling possibilities are available (see for example (Zollman 2013)), but it will be useful to consider the basics of signalling theory, and a basic taxonomy of the models available in the context of religious signalling.

4.1. Signals and signalling systems

Mirroring the heterogeneity of the literature on religious signalling, signalling theory outside of the religious context has been variously developed in such contexts as biological mate selection (Zahavi 1975), ecology (Hebets et al. 2016), and in formal modelling and philosophy of biology (Grafen 1990a; S. M. Huttegger and Zollman 2013; Skyrms 2010). The canonical starting point for current signalling theory is (Maynard-Smith and Harper 2003), which I draw on here for sake of precisifying some of the relevant terms and concepts.

4.1.1. Signals and cues

Strictly speaking, any observable trait that is correlated with an unobservable one can be said to signify the underlying trait and thus be a 'signal' of it, in the loosest sense of the word. The first distinction to make though is between cues and signals proper. *Cues* are observable features of an organism which correlate with other features but did not originate because of that correlation. For Maynard Smith and Harper, the distinction between cues and signals is that signals are: "any act or structure which alters the behaviour of other organisms, which evolved because of that effect, and which is effective because the receiver's response has also evolved" (Maynard-Smith and Harper 2003, 3). Many scientists and philosophers of biology follow suit

here, though there is some divergence in emphasis on the co-evolution or mutual benefit of the sender and receiver mechanisms. For example, by the letter of the given definition, it not necessary that the evolution of the response behaviours was entirely in the context of these same signals and in response to them. This is perhaps because, for these authors, the intent was to be able to draw a distinction between genuine behavioural responses (such as shying away from a bellowed warning) and non-behavioural responses (falling over when pushed), and to then differentiate signals from cues, where the observable feature of the sender is accidental with respect to the information it provides to the receiver.

In other words, something that signifies is a *signal* if there is also some sort of communicative agency or teleology on the part of the sender/signifier (cues are observed, signals are both observed and sent). There is broad scope for realising that criterion: the sender and receiver strategies which constitute signalling relationships can be passive or behavioural, and the result of natural selection or other mechanisms.

Consider the following illustration. A newly introduced species of snake becomes established in an environment, where it predates on several species of frogs, one of which is mildly toxic to the snake. The snakes rapidly learn to avoid that species, identifying it via a characteristic patch of colouring on its back. In that case, the colour patch is a cue. But if the more prominently coloured members of the frog population are predated less frequently, we might expect future populations have more salient colouring. The colouring would then be more signal-like, because (with respect to its evolutionary trajectory) it is now an adaptive response to the receiver's (in-turn adaptive) discrimination on the basis of it. The signal and response are now co-evolving, as the predator avoiding this particular prey is in the fitness interests of both parties. Coordinating on the signal facilitates that. Other signalling modalities are also possible of course: if the frog's distinctive colouring was on its webbing (for example) then it might evolve an active signalling display where it raises its webbing when sighting the predator. Passive traits and behavioural traits both count as signals in this regard, but it is the co-evolutionary relationship between sender and receiver 'strategies' that is important. Once the receiver begins conditionalizing their behaviour on a proto signal, it is in the interest of both parties to improve their coordination around it.

This illustrates some of the key points to be carried forward. The first is that while conceptual distinctions such as between cues and signals are clear-cut, real-world cases can admit of

degrees. Cues can gradually evolve into signals with borderline and hybrid cue-signal cases possible, during evolutionary transitions or otherwise. The second is that telling the difference between cues and signals might require knowledge of recent evolutionary history: is some signifying trait a happy accident, or has it come about (in its current form) for the very purpose of advertising what it signifies? Additionally, this is more than just a matter of stipulative terminology or semantics. Joint evolutionary histories (or at least trajectories) are important from the point of view of explanation: a one-sided, un-evolved cue-observer relationship does not by itself explain the existence of the signifier trait. *Explanatory* signalling therefore requires convergent teleological backstories for both senders and receivers, be they provided by natural selection or some other mechanism.

4.1.2. Manipulation, and semantics vs substance

Cues are where the sender doesn't participate in signalling relationship, what of the inverse possibility, where teleology is lacking on the receiver side instead? This is where terminology become less uniform. Consider the *Ophrys speculum* orchid, which attracts male *Campsoscolia 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 et al. 2003).

Following (Owren, Rendall, and Ryan 2010) and earlier precedents (Richard Dawkins and Krebs 1978; Krebs and Dawkins 1984), cases like this are described as influence or *manipulation* rather than communication, information, or signalling proper. Other writers in this tradition (the tradition of defining signalling with respect to evolved function) make the further definitional stipulation that the sender's and receiver's interests *must* be in accordance, such that blatant manipulation of receivers by senders falls short of signalling or communication proper; see for example (Scott-Phillips 2008). Purists such as Scott-Phillips would balk at the idea that 'senders' who have evolved a way to manipulate pre-existing responses of 'receivers' might count as taking part in communication or transmitting information per se.

Table 4-1 contrasts the relevant options here in a simple 2x2 matrix, according to whether the sender and receiver mechanisms evolved in response to one another, or whether they are (in that sense at least) accidental. In the case where sender and receiver mechanisms co-evolved

(SE-RE), we have uncontroversial signalling, but in SA-RE (where the sender mechanism did not evolve in order to inform the receiver) the result is a cue rather than a signal.

Table 4-1 Evolutionary classification of cues, signals, and manipulation[41]

		Receiver mechanism	
		Evolved in response (RE)	*Not evolved in response (RA)*
Sender mechanism	*Evolved in response (SE)*	Signal	Manipulative (signal?)
	Not evolved in response (SA)	Cue	?

The contentious cell here is SE-RA. Maynard Smith and Harper's position is ambiguous. They describe what the sender is doing as a manipulative signal (the sender has evolved a mechanism to exploit a pre-existing mechanism of the receiver), which is unlikely to be evolutionarily stable unless behaving according to that mechanism is likely to be beneficial for the receiver as well. And while they acknowledge that an evolved signal can be seen as manipulative (if there is any less-than-perfect alignment of interests), they appear comfortable using the term signal without 'manipulative' being some sort of negating term. A similar usage, acknowledging Scott-Phillips but proceeding in spite of it, is seen in (Ruxton and Schaefer 2011).

Yet another definitional approach, offered in a recent primer on animal signals in Current Biology (Laidre and Johnstone 2013), stipulates four conditions for signalling that are each designed for applicability to empirical study. On this account, signals are:

(1) acts or structures produced by signallers, which
(2) evolved for the purpose of conveying information to recipients, such that:
(3) the information elicits a response in recipients, and
(4) the response results in fitness consequences that, on average, are positive for both the signaller and the recipient.

This is similar to the (in principle) lop-sided origin requirement of Maynard-Smith and Harper, but with an explicit additional condition (4) that definitionally builds-in the future-looking

[41] Similar classification tables are used in (Scott-Phillips 2008) and (Diggle et al. 2007).

alignment of interests. This then makes explicit the ways in which evolutionary definitions of signalling can come apart: they can apply evolutionary conditions to sender or receiver mechanisms (or both), and those conditions might require a past evolutionary history or a current pro-fitness effect (or both). While Maynard-Smith and Harper and Laidre and Johnstone mix and match these requirements, the Scott-Phillips camp would insist on all being applied to both sender and receiver: like cues, manipulative signals are not signals proper. In any case, we can also recognise here an awkward mix-match of fitness consequence and evolutionary history criteria being applied.

The awkwardly rigid definitions, awkwardly inconsistent definitions, and the general lack of unity in the biological literature are relevant here because we see something similar in the religious signalling literature. The terms ‘signal’ and ‘signalling’ show a similarly loose usage: both (Cronk 1994) and (Sosis and Alcorta 2003) for example discuss the likelihood and mechanics of manipulation *via* religious signalling, mirroring the language of Maynard Smith and Harper. This is despite of (or perhaps because of) a general awareness of the precedent of Dawkins and Krebs on signals as manipulation. Reading off what exactly is meant by religious signalling is therefore potentially problematic, as is retrofitting these intentions into a more strictly defined conceptual framework.

The key issue in dispute here is whether manipulative signals are signals, i.e. whether signals require co-evolution or common interest. I do not think much hangs on this determination, but we should be aware of both the potential for equivocation if it is not made and the conceptual trade-offs if it is. For example, one consequence of a tight, Scott-Phillips-style evolutionary definition is that it becomes difficult to find many cases of dishonest signalling (though dishonest ‘signalling’ abounds). Indeed, the implication that signals and communication are honest *by definition* sits poorly with natural language. False signals are still (intuitively) signals. While it might be more plausible that *communication* is not actually taking place in the flower-wasp case, what is it that the male wasp receives from the flower (and conditionalizes their behaviour upon), if it is not a signal? This is certainly the parlance of the empirical scientists who study the phenomena (such as Ayasse et al.), who are happy with the term ‘chemical signals’ to describe what it is *Ophrys speculum* produces. It is of course possible to insist that the empirical scientists are speaking figuratively or improperly when they talk like this. Even if there were theoretical merits to this (which remain mysterious to me), terminological revisionism is not pragmatically advisable. Most obviously, it is unlikely to be taken up by the

empiricists, likely resulting in a semantic mismatch between them and the revisionist theorists with all the attendant risks of talking past one other. And yet there is something genuinely useful and categorically neat about the strict conceptual partitioning summarised in table 4-1.

I suggest that the terminological incongruence here is evidence of different ways of precisifying 'signal' and 'communication' based on local disciplinary convenience and divergent demands for explanatory utility and/or application. Different clusters of users, with distinct clusters of uses and theoretical approaches will precisify slightly different concepts of signalling and communication as the most appropriate (i.e. from that theoretical standpoint or within those particular ecosystems of inquiry). For example, Scott-Phillips is specifically focused on defining biological communication in the abstract, and all his examples are of systems evolved via genetic inheritance. Trivially though, communications technology does not come about that way. Less trivially, rough and ready notions of 'signal' and 'communication' are obviously acceptable for some field biologists. In the context of religious signalling we should also probably allow the possibility of selection operating on cultural inheritance, or via non-selectionist mechanisms which merely approximate the effects of selection (i.e. sufficient for signalling models to be reasonably applied). An apparent dispute about how to define communication (or just biological communication) can be less a scientific or philosophical matter as one of sociological/disciplinary or contextual mismatch, and there are a standard set of options for resolving such disputes. If the primary concern is evolutionary theory and the evolution of stable communicative traditions in the abstract, it arguably makes more sense to 'carve the joints' of the world with signal-talk anchored to common interest/co-evolution. At least, it does so more in that scientific context than it does if one is a field biologist trying to make sense of the strange behaviours and morphologies of various pairings of flower and insect. It is not clear to me that either disciplinary focus has first dibs on the terms, and we should perhaps be content with a bit of terminological polysemy[42].

I would also argue that (Richard Dawkins and Krebs 1978; Krebs and Dawkins 1984) should be identified as sources of some of the terminological and conceptual unevenness here. In these two influential chapters (from two early editions of a seminal behavioural ecology handbook) the authors stress that any signal (co-adaptive or not) can be seen as a manipulation of the

[42] See (C. Brusse 2016) for an extended case study and discussion of terminological disputes in a mixed-specialisation context.

receiver. By one idiom of natural language this is true, even in a co-evolutionary, co-adaptive signalling relationship the sender is trying to get the receiver to do something. But in such a relationship it is equally true that the receiver exploits the sender's willingness to be so open. Both sides are in it for themselves. So, while it is appropriate to apply evolutionary de-romanticism to debunk any idea of naïve, 'kumbaya'-like communication for the common good in the natural world, it is a mistake to see the overall signalling relationship as entirely (or even predominantly) constituted by what the *sender* is doing, in some sort of active-passive dichotomy. This should be kept in mind, even (or especially) in cases which strike us as blatantly exploitative. For example, think of a low-ranked male chimpanzee surreptitiously displaying an erect penis to a female: the male is certainly relying on pre-evolved cue-response mechanisms to further his own ends, but there is no reason to suppose that the female will not respond according to her own interests (Nadler and Bartlett 1997; Van Lawick-Goodall 1968). Senders generate the signal, receivers trigger the payoff, but in a *co-adapted* signalling relationship neither can be said to be entirely in the driver's seat, or entirely passive. Each takes advantage of the other, so over-interpreting the signals-as-manipulation description is unwise.

What this means for current purposes is that we should be cautious of both received terminology and our own intuitions. Regarding manipulation, what is important is that manipulative signals and co-adaptive signalling are *not* of an explanatory kind, any more than cues and signals are of a kind. We need not be prescriptive with language, but we should be wary of it. In loose, functional sense of the term, the orchid does 'signal' that it is a female wasp, just as cue trait in some sense 'signals' for what it is recognised as signifying. For sake of completeness, we can also imagine filling the fourth cell of table 4-1 with examples of introduced species who just happen to have complementary strategies and behaviours (and/or repurposed parts from two kinds of communication devices which just happen to work together). What passes between them might be signals in the broad sense, but the relationship between sender and receiver is not a good candidate for explaining the relevant behaviours of either. Neither does any contingent alignment of interests (at least not initially). For the fitness value of signalling to explain it and the behaviours which realise it, senders and receivers must

share a recent co-evolutionary history[43], in the context of which their relevant behaviours and strategies can be seen as adaptive.

4.1.3. Signalling systems and formal modelling of religious signalling

To sidestep the semantics of 'signal', I will be following David Lewis (Lewis 1969; Skyrms 2010) in describing these mutually adaptive signalling relationships as grounded in game-theoretic *signalling systems*. A signalling system is a solution to a strategic situation where the sender conditionalizes their behaviour on the underlying trait of interest, which the receiver observes and in return conditionalizes their own actions upon. Given a sufficient alignment of interests, such conditional strategy combinations can constitute Nash equilibria: where neither party has incentive to alter their strategy. If further conditions hold, they can be evolutionarily stable, or even guaranteed to emerge under standard models of evolutionary change (Simon M. Huttegger 2007; Pawlowitsch 2008). Lewis's notion of signalling systems, and the formal modelling literature that has subsequently followed from it, therefore nicely underwrites the sort of co-evolutionary signalling relationship we are interested in.

The formal modelling of relevant signalling forms will be considered in depth in chapter 6. But this formal literature also offers a few suggestions and caveats which should be noted in passing. One of these relates to an obvious worry that I have been dancing around: is natural selection really a plausible vehicle for the evolution of *religious* signalling? We can certainly understand the evolution of the colouring of a toxic prey animal and recognition/avoidance behaviour of a predator as two sides of a biologically evolving signalling system. Communicative strategies in this case are plausibly 'in the genes' and less communicative strategies can be fatal. But the religious behaviours that would be relevant to a signalling hypothesis are much less plausibly, robustly wedded to biological lineage, vertical reproduction of phenotypes, and natural selection. Religious ritual is something that one learns, perhaps horizontally or obliquely from peers and other non-parents, and if it is a communicative strategy then the mechanisms which select for it are less obvious. There is therefore an element of mismatch between model and target system, if the model assumes evolution by natural selection.

---

[43] This appeal to 'recent' evolutionary history of course requires a further story, but such stories are available from other contexts in philosophy of biology. See for example (Griffiths 1993) and (Godfrey-Smith 1994) with respect to etiological theories of proper functions.

Of course, genic natural selection is not the only game in town, as seen in chapter 1, and I will not delve deeper into the empirical case for this. But the formal literature also offers support here because, despite the biological examples it is often couched in, the signalling system framework extremely flexible. In evolutionary game theory (Maynard Smith 1982a), natural selection via differential fitness is typically modelled using the *replicator dynamics*, with the relative expected payoffs of different strategies interpreted as relative fitnesses. Each strategy in a population of senders and receivers will have an expected payoff that depends (primarily) on the proportion of potential interaction partners whose strategies complement, reward, or exploit that strategy. Over multiple iterations, the replicator dynamics adjusts those proportions up or down depending on how expected strategy payoffs compare with their competitors, in a way that mirrors idealised biological evolution. But this is only one among many evolutionary dynamics (Cressman and Tao 2014; Nowak 2006; Sandholm 2011), and various kinds of payoff-sensitive reinforcement dynamics and learning dynamics (such as "imitate success") have been shown produce similar results with respect to signalling systems and the stability of signalling behaviour (Skyrms 2010). These dynamics are designed to mirror non-reproductive processes of behavioural phenotypic change, e.g. the aping of successful behaviours or other cultural-evolutionary transmission mechanisms. And the replicator dynamics too are sometimes used agnostically: without making any assumptions about what sort of mechanisms might be approximating selection processes. This therefore invites interpretations of sender-receiver strategies evolving and converging over more flexible timescales via cultural or rational processes; what many of them have in common though is some form of sensitivity to success or failure with respect to differential payoffs[44]. The upshot for current purposes is that 'payoff-driven' or 'adaptive' notions need not be too specifically applied at this point: we should not assume that natural selection is the only adaptive mechanism which might reify the signalling system framework.

With suitable caveats in place then, the more detailed characterisation I am advocating is as follows: the RST involves positing signalling systems such that religious signalling came about (or became locked-in) because of co-adaptive behavioural traits on the part of the participants

[44] It should be stressed that the use of evolutionary dynamics in the formal modelling literature is *not* typically to model putative evolutionary narratives. Rather, evolutionary dynamics are used in simulations are to analyse the parameter spaces of various models, and determine the evolutionary stability and basins of attraction for communicative strategies and their equilibria (see also chapter 7).

and observers of religious ritual. The evolution of religious signalling on this view is payoff-driven, because honest signals of specific social norm adherence (prosociality) and sensitivity to them were both rewarded by positive assortment with normatively compatible conspecifics. More robust signalling allows greater positive assortment of altruistic types, with a corresponding increase in the cooperative dividend, in turn further incentivising efficient signalling systems. Of course, this dividend also incentivises cheating – all ecosystems have their parasites and to survive them there must be safeguards. As before, bullies will attempt to simply take what they want and will need to be resisted (so cooperation being signalling-mediated matters little in this regard). But free riding in this context becomes subject to the ability to convincingly send manipulative signals: i.e. lie. As is obvious, the value of signalling-mediated cooperation is therefore dependant on honesty-biasing mechanisms which stymie that. Religious signalling must remain reliably correlated with cooperative value if a signalling system is to be stable enough to in turn stabilise cooperation.

4.2. Honesty in fakeable and unfakeable signals

There are two canonical subdivisions of signalling, each with their own honesty-biasing mechanism. For the most part I will be calling these *fakeable* signals and *unfakeable* signalling; though the nomenclature (and spelling) is not uniform. For example, unfakeable signals are sometimes called *hard to fake* (because fakeability admits of degrees), *index* (though 'index' is sometimes used for any correlation between signal and underlying trait), *constrained* (Searcy and Nowicki 2005) or *performance* signals (Hurd and Enquist 2005). Hurd and Enquist also contrast these with *strategic* (i.e. fakeable) signals, contrast both against 'out of equilibrium' signals (such as manipulative signals or other signals not part of signalling systems), but also use 'index signal' as a sub-class of their performance signals. This diversity of terms is unfortunate, with the inconsistent use of terms like 'hard to fake' sometimes figuring in the conflation of signalling types (e.g. see chapter 5). The goal of the rest of this chapter is to tease the relevant distinctions apart in the context of religious signalling, without falling into such traps.

4.2.1. Unfakeable signals and the fakeable-unfakeable distinction

Unfakeable signals are intuitive: they reliably convey information about an underlying state or trait by using a method that is difficult or impossible to replicate in the absence of that underlying state. Incentives and disincentives are not part of the causal picture. Maynard-Smith

& Harper's example is the roar of a stag in rut, the qualities of which (pitch, timbre, etc) reliably index its size and strength, thereby providing information for other stags to conditionalize upon. It is therefore easy to see how this signalling system evolved: weak stags can avoid picking fights with strong stags, who by scaring them off can save their strength for genuine rivals. When the generative connection between signified and signifier traits is not diaphanous like this (or where alternative generation methods are exploitable), the signals are more open to being sent by agents who lack the underlying qualities of interest, i.e. they are fakeable.

Intuitive as it is, the distinction is not entirely straightforward. First, and most obviously, the distinction can admit of degrees. A sender might be able to 'fake it', but only with an imperfect success rate. Indeed, fakeable and unfakeable signalling can be brought under formal generalisations which have them admitting of degrees, though they are not simply paraphrases of each other at the formal level (Simon M. Huttegger, Bruner, and Zollman 2015)). Proper application of fakeable and unfakeable models to real-world cases might also be sensitive to evolutionary timescale and ecological context. For example, the colouring signal in the frog case might eventually be mimicked by another species to avoid predation. Being mimicked in isolation would mean that the signal is (to that extent) fakeable. But perhaps the colour is derived from the chemical toxin itself, and the only (or most accessible) evolutionary pathway to mimicry for relevant frog species is to evolve similar toxicity – and this would effectively be the parallel evolution of the same, genuine, hard-to-fake signal (at least as far as the snake is concerned).

The distinction itself has also been challenged at the theoretical level, starting with (Grafen 1990a). Grafen argued that when viewed over timescales of developmental investment or evolutionary strategy, hard-to-fake signals are just special cases of fakeable signals. What makes them *look* unfakeable are developmental costs and sub-optimal side-effects which prevent the evolution of hoax signalling mechanisms but do so via prohibitive *disincentive* rather than nomological constraint. For example, such considerations prevent the developmental investment by smaller stags in some sort of dedicated 'deep-roaring' throat structure, but the in-principle possibility of such a structure renders the roar fakeable (J. M. Biernaskie, Grafen, and Perry 2014; Searcy and Nowicki 2005, 216). On this view, unfakeable signals are fakeable after all (viewed from a suitably long timescale), they are just *not* being faked because the costs of doing so would be too high.

Grafen's argument may be theoretically sound, but it can be side-stepped. If suitable adaptive pathways are available over relevant timescales, then it might be fair to say that the classification of any given signalling system as fakeable or unfakeable will be a modelling decision, based on the scope and operational timescale of the intended target system so described, and the relevance of various counterfactuals. But modelling decisions can be defended and we can use explanatory context to carve off some pragmatic distance from the in-principle reducibility of unfakeable to fakeable. For example, if the signalling trait we are interested in is *known* to have evolved on a shorter timescale than is plausible for some deep developmental change to make them relevantly fakeable, then that is not actually a relevant explanatory alternative. In this case we can treat the unfakeability-fixing constraint as given. For purposes of religious signalling, such constraints might include generatively entrenched cognitive features, or susceptibilities to certain types of responses: basal emotional responses to stimuli, or the use of pain and sensory overload. It is not reasonable to expect these to be shifted much by the evolutionary pressures of a few hundred or thousands of years of exposure to a religious signalling tradition. Likewise, we might expect the general human ability to lie about one's commitment or devotion to gradually improve over time, given enough exposure to fitness-impacting signalling systems that leverage our poor acting ability. But the rate of improvement would be glacial compared to the much shorter-term evolution of specific, comparatively ephemeral religious cultural forms (as candidates for signalling). There is more to be said, but for current purposes (and with these caveats in mind) I will be assuming that the distinction is good enough to work with.

4.2.2. Fakeable signals and costly signalling

What of keeping fakeable signals honest? Without going too far into the game-theoretic reasoning for the moment, Lewis's original work was on systems of costless fakeable signals (which he treated as an idealisation of language) in games of common interest – where the defining payoff structure for the signalling game is set so that sender and receiver should prefer identical outcomes. Signalling systems in this context are solutions to coordination problems. Under these conditions, 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. But these 'cheap talk' equilibria become far less stable as interests start to diverge (Godfrey-Smith and Martínez 2013; Martínez and Godfrey-Smith 2016). 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 the same incentive to use any "I'm a high type" signal that evolves.

Enter the misleading elephant(s) in the room: costly signalling and the handicap principle. The basis of costly signalling or *handicap* signalling systems (Grafen 1990a; Zahavi 1975; Zahavi and Zahavi 1997) is that the incentive to lie can be mitigated and reversed if the costs of sending a signal are such that lying is no longer worth it. This was implicit in the discussion of Grafen, and costly signalling and the handicap principle is often alluded to in the religious signalling literature (Sosis and Alcorta 2003; Sosis 2003; Bulbulia 2004a; 2004b; Bulbulia and Sosis 2011; C. Alcorta 2017).

Zahavi's initial focus in (Zahavi 1975) was to apply the handicap principle to mate choice, and it remains an illustrative example. Imagine a population of females who want to be able to discriminate between high quality and low-quality males, but males of every type just want to be chosen, i.e. to be treated as high types for mating. Simple, cost-free self-reporting by the males will therefore be as good as useless: low types and genuine high types alike will have just as much interest in signalling that they are high quality mates. It is therefore in the interest of the selectors (the females in this example) and the high-type senders (the males) to find a way to alter the calculus of the low-type sender so that it is no longer in the interest of low-types to self-report as high. This is the evolutionary rationale which leads to the handicap principle, because one way of altering the calculus for the low types is for high types and receivers to display and recognise a 'marker of quality' which self-handicaps the selected sex/population, such that adopting the marker is simply not worth it for the low types – while still worthwhile for high types. Such a collusion against the low types would be a stable, self-sustaining signalling system and (though this is a separate question) to a first approximation it might be reasonable to expect that mutations be selected for if they optimise the behavioural phenotypes toward such a signalling system.

The more famous 'poster child' for the handicap principle (and costly signalling theory in general) is the stotting behaviour of various species of Gazelles (including Thompson's Gazelles), when approached by a leopard or other pursuit predator. Stotting does not make much intrinsic sense given the Gazelle's need to evade if pursued, but it does make sense as an optional handicap that the predator has co-evolved to recognise. The interpretation is that fast, fit Gazelles (high type senders) who stand a better chance of outrunning the predator (receiver)

can afford to waste some time and energy to broadcast this (costly signal) to the predator, who can then redirect their own energies toward running down the more vulnerable, slower Gazelles for whom stotting would be too costly (low type senders), to the mutual benefit of the predator and the high types. In its most basic form, a costly signalling theory of *religion* would assign these four abstract roles to: cooperative or otherwise socially valuable individuals (high types), the general community (receivers) who must decide which individuals to cooperate with and/or permit entry, religious ritual participation (costly signal/handicap), and less socially valuable individuals with lower levels of commitment (low types).

4.3. The problem with 'costly signalling'

Famous as it is, the handicap principle has a convoluted history of conception and application (Grose 2011), and some of this is highly relevant to RST. As described Grose, and by (Pomiankowski and Iwasa 1998) in review of the far more ambitious (Zahavi and Zahavi 1997), the handicap principle was initially treated with caution, and its claims to universality and/or relevance remain controversial[45]. Indeed, the Zahavian program took a while to clarify. One obvious early objection in (Maynard Smith 1976) to the initial description of it (Zahavi 1975) was that while it is in the interest of females to discern high-type males, it is not necessarily in their interest to mate with self-handicapping males if they are likely to pass this handicap trait on to their offspring. A similar argument is made by (J. W. F. Davis and O'Donald 1976), which also focuses on the sexual selection case and the lack of clear advantage for the costly signal in the long-term. These objections were largely addressed in (Zahavi 1977), which stresses the need for *differentially* costly signals: many of the initial misgivings were due to a tacit assumption that signal cost would be the same regardless of the sender (i.e. no greater opportunity cost or other form of cost for low-types), and this would indeed fail to give the signals any evolutionary advantage. The original choice and focus on mating display examples also needlessly framed costly signalling in competition with the established and more popular explanations of sexual selection, such as the Fisherian 'runaway' process (Fisher 1930).

As Grose argues, the primary driver of the development of costly signalling theory has been formal modelling. Zahavi and co-authors did not offer a formal model of the principle and it

[45] Pomiankowski and Iwasa point out that empirical verification of the expansive claims is lacking, and describe the book as being "stuffed full of the Zahavi's endless speculations about the value of different signals" and "an odd mixture of home truths and wacky fantasies".

was not until much later that compelling models were put forward in the biological literature which captured and formally validated the handicap principle (Grafen 1990a; 1990b; Maynard Smith 1991; Johnstone and Grafen 1992). Early critical models in biology such as Maynard Smith's failed to recover the promised benefits of the handicap principle, and without ongoing dialogue at the level of formal modelling, progress stalled. This was despite formal models of costly signalling being widely used in economics (Riley 2001; Spence 1973; 2002). Though academic 'siloing' might be in part to blame, in part it is probably also due to the difficulty of pinning down exactly what is meant by a handicap or a costly signal in a biological or cultural (or at least non-monetary) context. Fitness costs are hard to infer, and proxies for them can be obscure or deceptive. As Maynard Smith later wrote with Harper:

> "given the models, it is fairly easy to spot the difference in the assumptions responsible for the different conclusions, but without the formal models it would be hard to do so. Given a formal model, it is possible to define terms such as handicap, index, cost, cue, and so on in a relatively unambiguous way" (Maynard-Smith and Harper 2003, 3).

The corollary to this though is that the reverse-process of interpreting these formal models onto real-world, putative signalling phenomena is non-trivial. In the context of religious signalling this should give us pause; in the biological and cultural domain there is a challenging disconnect to be bridged between signalling model and putative signal 'in the world'. This is the subject of 4.3.3 and much of chapter 6.

It should be noted in passing though that religion is in no way the sole human application of signalling theory. More direct applications of signalling to altruism include proposals that spiteful punishment is a costly signal of prosocial quality (Jordan et al. 2016; Jordan and Rand 2017). In interpreting costly behaviours as signals RST has close parallels in other areas of social science, one striking example being attempts to explain suicide bombing as an adaptive, costly signal not for individuals (for obvious reasons) but with ideological groups playing the signalling role (Lapan and Sandler 1993; Hoffman and McCormick 2004; Pape 2006; Zahedzadeh 2017). With regard to personal social psychology, everything from grief, to smiles and apologies have been interpreted as a costly signals of prosocial traits (Ohtsubo and Watanabe 2009; Winegard et al. 2014; Centorrino et al. 2015; Rosenstock and O'Connor 2018), and altruism itself has even been proposed as a costly signal of intelligence (Millet and Dewitte 2007). However (Grose 2011) makes a convincing case that some such accounts are worrying

imprecise with regards to the way that costs are inferred and the work that signal costs are supposed to do; especially with regard to misapplications of the differential cost condition, to which I now turn.

4.3.1. The differential objectivity challenge

There are several conceptual distinctions and methodological lessons and to be extracted from the historical and broader academic context of costly signalling. The most obvious point is that it is not enough for costly signalling systems to exist that 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: i) the average cost (c_l) of an "I'm a high type" signal by a *low*-type must be greater than the average benefit of them being treated as a high type (b_l), while ii) the average cost of the honest version of that signal (c_h, for a genuine high-type) must be less than the relevant benefit (b_h), i.e.:

$$c_h < b_h, \qquad c_l > b_l \tag{4.1}$$

This joint inequality condition can be simplified to a single, intuitive inequality for the standard differentially costly signalling model (assuming benefits are the same, i.e. $b_l = b_h = b$):

$$c_h < b < c_l \tag{4.2}$$

For receivers (and high-type senders), it is therefore strategic to coordinate on a signalling mechanism which imposes differential costs on the sender. Absent these conditions, and there is no such incentive, and no hope for a signalling system (as long as the signals are fakeable).

This has not been entirely ignored in the religious signalling literature. For example, in laying out the CREDs theory as an alternative to costly religious signalling, Henrich points out 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 similar worry acknowledging the differential costs condition was previously voiced in (Sosis 2003). Indeed, hours spent sitting on a pew take the same time away from more profitable activities, regardless of sincerity. And if tithes, arduous rituals, and other demands to secure and/or maintain one's good standing in the religious community are (by hypothesis) *tests* of commitment, 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 influence behaviour accordingly, with 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 (by breaking the adaptivity of the inner co-evolutionary relationship). I will return to this in next chapter, as it is the approach favoured in (Sosis 2003). But for the moment we can note that if the driver of signal evolution is perception rather than relative fitness (or some other kind of objective cost-benefit consideration) then this begins to look more like a cognitive bias theory than RST proper. In any case, there appears to be 'differential objectivity' challenge for costly signalling in the religious case.

What appears to be less appreciated in the religious signalling literature (and perhaps the broader 'applications' literature in general) is that the differential objectivity challenge requires that we assume, as condition [4.2] does, that the benefits of being treated as a high type are always the same, i.e. $b_l = b_h = b$. If we relax that assumption, then we can easily accommodate equal signal costs $c_l = c_h = c$, providing a different condition is satisfied:

$$b_h > c > b_l \qquad [4.3]$$

This 'differential benefit' signalling is often mentioned as a corollary to differential cost signalling[46] (Bulbulia and Sosis 2011; Johnstone 1997; Maynard Smith 1991; Murray and Moore 2009). As should be expected, its evolutionary dynamics mirror that of differential cost signalling (Zollman, Bergstrom, and Huttegger 2013). However, it is arguably less well explored in terms of application to real-world cases.

In the following chapter I will argue that hard-to-fake signalling and differential-benefit fakeable signalling models both have theoretically plausible interpretations for RST purposes, with more exotic differential-cost models potentially available as well. But these modelling strategies need to be carefully disentangled from the (unfortunately) more familiar notions of handicap and costly signalling.

For the moment though, we can flag that what I am calling the differential objectivity challenge is vulnerable both to being under-appreciated and being over-emphasised. At the level of signalling theory alone, there are two clear reasons why the differential objectivity challenge might be less than fatal, because we do *not* need to understand RST as a differentially costly

[46] In another example terminological inconsistency, the term 'handicap' is sometimes applied by some only to signal costs (excluding differential benefit signalling), but in other usage subsumes differential benefit signalling. Of course, for Grafen et al, the term ultimately subsumes index/unfakeable signalling as well.

signalling hypothesis. RST-theorists could emphasise instead cognitively rich, hard-to-fake signalling; and several have done so (Bulbulia 2013; Bulbulia and Sosis 2011; J. H. Shaver and Bulbulia 2016). They could also retain the differential cost-benefit incentivisation framework for fakeable signals, but instead explore options for differential benefit signalling systems.

4.3.2. The function of cost and benefit

The need for differential costs and/or benefits for fakeable signalling systems (rather than just costly signals) is not the only reason why the terms 'costly signalling' and 'costly signalling theory of religion' are misleading, and not the only source of confusion within the literature[47]. For example, it is not even the case that costly signals imply fakeable signalling systems. Unfakeable/indexical signals can have costs associated with producing them, usually dubbed 'efficacy costs' in contrast to the 'strategic' costs of signals in conditions [4.1] and [4.2] (Lachmann, Számadó, and Bergstrom 2001)[48]. The stag's roar for example might be energetically expensive or might risk the attention of predators, perhaps translating into non-trivial fitness cost. But assuming the relative costs and risk for high and low types are similar, they could not be *strategic* in the sense of being differentially incentivising (assuming similar benefits also). Flat signal costs like this are a burden to all. One upshot of this is that available evolutionary pathways to signal more cost-effectively will be favoured (this marks out one evolutionary difference from strategic cost signalling). In any case, if no cost-free signalling methods are available (and the benefits of a favourable response make them worthwhile), index signals will be costly, and so we can't make the inference from costly signal to handicap signal. Breaking the false bi-conditional from the other direction (following (Számadó 2011)), a costly/handicap signalling system need not be producing observably costly signals either. One intuitive way this can happen (i.e. with condition [4.2] being satisfied) is when c_h is zero (i.e. only low types pay a significant signal cost $c_l > b_l$), and once the population has had time to reach equilibrium – because at equilibrium the low types shouldn't be sending their benefit-exceeding costly signals anyway. The variables in [4.2] and [4.3] can have any cardinality (positive or negative) as long as they maintain their ordinal ranking, making it relatively trivial to define fakeable signalling systems with costs and benefits that might seem counterintuitive.

---

[47] I have been cagey about talking about costly signals (and examples of them) in the earlier exposition of core RST and general signalling theory, largely because of the scope for distraction that these confusions provide.

[48] This is in the biology-influenced signalling literature at least. The generally overlooked parallel in the economics literature are costs associated with "certification" (J. P. Bruner 2015a).

In short, the presence of costs is neither necessary nor sufficient for the costly signalling or the 'handicap principle' (or differential cost-benefit signalling in general) to be in play.

This is even more obvious when we consider costs which attach to displays or behaviours that are not genuine signals (in the signalling system sense). What looks like a costly signal might not be part of a signalling system at all. The CREDs mechanism is one clean example of an alternative explanatory scenario, where what looks like a costly signal of some deeper, hidden social quality (genuinely valuable to the receiver) is just a costly display of sincerity with respect to the first-order observable content[49]. Costs might also be evolutionarily functional as a way of psychologically influencing those subject to them (rather than those observing): binding participants in a common experience of hardship (e.g. Whitehouse's 'social technology' mechanism), or just via sunk cost fallacy. Ritual costs might also constitute a form of perverse maladaptive subjugation, put in place to demonstrate the power of the leadership hierarchy (or for the entertainment of existing group members, and so-forth). Of course, depending on how costs are paid, they might instead be an extortive transfer of resources. Straightforwardly interpreting observed costs as signal costs (in the relevant sense) is inherently problematic.

4.3.3. The currency of cost and benefit

So, when *can* we say that behavioural costs are signalling costs (of one type or another)? This depends on the degree to which we can measure or estimate them and compare them to the other costs and benefits the putative signalling system, in at least some approximation to an overall payoff structure. But even the general idea of a cost can be dangerously vague, given the precise theoretical/formal role it must play in signalling system explanations.

For comparison, the metabolic expenditures we can observe with signalling among simple microbial populations are reasonably interpretable as fitness costs (Birch 2017). Biological life at this level skates very close to the margins, and fuel for living and reproducing is by far the most important battlefield for evolutionary competition. But other intuitively 'costly' signals and proxies for fitness beyond this level become increasingly questionable. Energetic expenditures like a stag's roar, or painful extreme rituals might not translate into significant

[49] It is valuable to partner up with someone who signals themselves to be a true believer, so long as sending the signal (and/or being a true believer) correlates with partner quality. It does not pay if the signal only correlates with being a true believer.

fitness costs at all. The worry is that the intuitive notion of a cost (when appraising a biological or cultural signalling candidate) is broader than game-theoretic 'cost'. Cost-talk must therefore be sensitive to what (Grose 2011) calls the *currency* of the cost: posited costs and benefits are all supposed to combine into the payoff structure of a signalling game (and conditions 4.1 – 4.3), and therefore be i) commensurable, ii) additive, and iii) significant relative to each other. This is potentially problematic for at least some interpretations of costly signalling in religion, for example it is not clear (at least to me) that there is any objective exchange rate between the psychological 'costs' of painful extreme rituals and the supposed economic benefits of improved partnership opportunities.

There are two related requirements to underline here. First, any appeal to game-theoretic solution concepts carries with it an obligation: any costs and benefits which are plausible to attribute to the various outcomes must be *relevantly commensurate* with one another. For each agent whose behaviour we wish to explain via costs and benefits, there must (in principle) be a common currency to cash them out, such that we can plausibly add and subtract them to produce an overall payoff that is supposed to drive the evolution of their behaviour.

Second, this currency should be paired with a plausible, and currency-relevant update mechanism for sender and receiver strategies, that (in principle) takes these payoffs – whichever manner of cost and benefit is used, it must be relevant to some plausible evolutionarily dynamic. That a Nash equilibrium exists explains nothing in and of itself, populations of agents must also have the means to gravitate toward that equilibrium. They must be *able* to update their strategies for adaptation to play any sort of explanatory role, otherwise there is no evolutionary explanation to be had.

None of this is to demand that theorists must explicitly quantify their theories and actually tote up numbers, or even that doing so according to their theories would be empirically tractable (though of course this would help). What it does mean is that it *might* be done in principle, and that the following be plausible: the variously identified costs and benefits really are being combined and compared in some way which approximates additivity, with respect to how agents respond to them over the evolutionary timescale posited.

This is of course another reason why modern economics has an obvious advantage: certain assumptions are made about economic rationality and the motivating value of monetary returns which fit evolutionary game theory like a glove. In the purest economic interpretation, payoffs

would be *utilities* (or even better, money), with costs and benefits representing modifications of overall utility. The appropriate update mechanism would be some kind of rational revision of strategies on observing relative profit or loss – perhaps linked to the relevant notion of utility (e.g. via revealed preferences). The 'evolutionary' timescale would therefore be rapid and not bound to reproduction, but this is still a recognisable, valid context for applying the general signalling system methodology, and indeed the sort of signalling models we are interested in (with some tweaks, including perhaps replacing the replicator dynamics with something more appropriate).

The requirements also seem to be satisfied for fitness costs and benefits in natural selection at a very simple level. As mentioned, microbial life has a simple currency of cost and benefit: they live, die and propagate according to the harvesting and expenditure of resources. Assuming some degree of strategy-heritability, a strategy with above-average fitness will have greater representation in the future population. Other update mechanism would be social learning, in which agents update their strategies either by observing the payoffs received by their fellows and/or by some other update strategy (e.g. conforming to majority practice).

Again, highlighting the currency problem in signalling theory provides no grounds for rejecting RSTs on the basis that they do not address it. But it does provide potential grounds for challenging them. If, on reflection, an application of signalling theory faces blatant interpretational issues regarding how costs and benefits are supposed to actually combine and drive the evolution of signalling systems, then it is reasonable to re-evaluate that application. Relying on signalling theory requires respecting the assumptions that validate it, and if they do not hold then the validity of that reliance is in question.

4.4. Why signalling theory matters

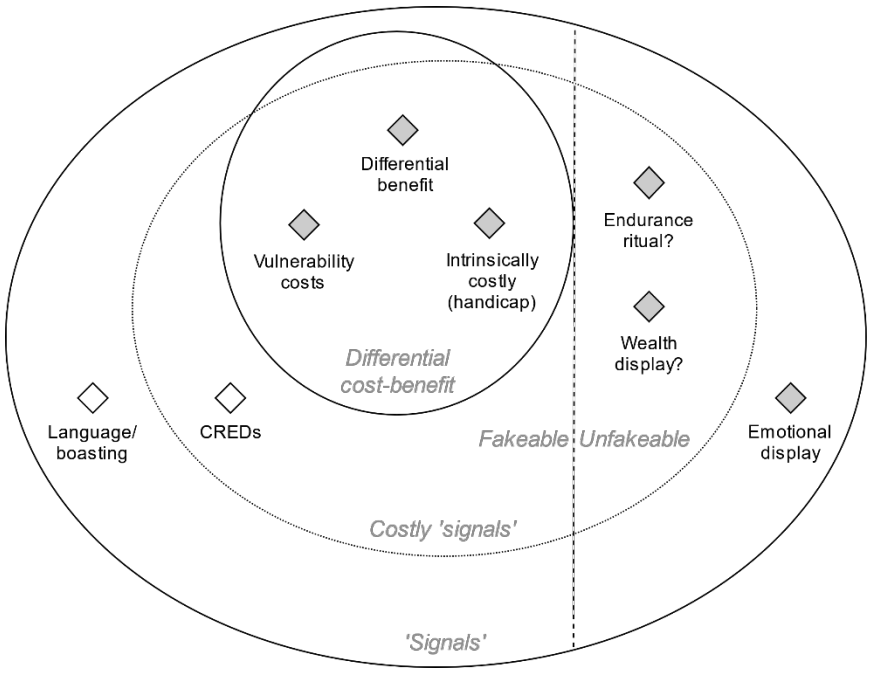
In the next chapter, in order to develop a more detailed picture of what signalling theory requires and how RST might meet it, I will be taking Maynard Smith and Harper's earlier recommendation to heart, and moving to actual models of signalling, formally defined. Maynard Smith's more famous quotation is also relevant: "Mathematics without natural history is sterile, but natural history without mathematics is muddled" (Maynard Smith 1982b). Before making things less muddled though, we can take stock of the various distinctions and caveats from this chapter, and what they signify.

4.4.1. Distinctions and difference-makers

We are now in a position to distinguish the various signalling models at a verbal level, to be formally fleshed out in chapter 5, in preparation for applying them to religious signalling in a more detailed manner (chapter 6) and testing them in ways motivated by those potential applications (chapters 7-onward).

Figure 4-1 provides a rough depiction of the space of signalling-like explanations I have described so far, in the context of partial conflict of interest (i.e. there being high and low type senders). The cost categories for signal-like behaviours is depicted in three more-or-less nested classes – 'signals', 'signals' with costs, and differential cost/benefit signalling, with costly 'signals' cross-cutting with the fakeable-unfakeable distinction. Shaded, labelled items are candidates for signalling systems under partial conflict of interest (i.e. signals proper). Remaining unshaded are costless fakeable signals (cheap talk), and flat cost mechanisms e.g. CREDs, which can be effective in the presence of appropriate behavioural biases but are do not constitute signalling systems in the adaptive, payoff-driven sense. Included under differential cost-benefit signalling are differential cost, differential benefit, and vulnerability cost signalling models (like differential cost, but where signal costs depend on the actions of the receiver, as well as the sender). Formal models of these will be considered in chapter 5.

Figure 4-1: Venn diagram classifying signalling models by costs/benefits and fakeability



As discussed, there is at least a terminological bias toward focusing on the distinction between costly and cheap signals. But this is much less of a natural kind distinction, as costly signals can be costly for different reasons and different ways. Focusing on signal cost alone is a red herring.

More important is distinguishing two classes of signalling system: i) differential cost- benefit systems and ii) systems based on unfakeable signals. This distinction is disputed by some (Grafen 1990a; Higham 2014), but is at least a *pragmatically* well-motivated one for our purposes. Signalling systems are contrasted here against other signalling-like phenomena such as CREDs, which propagate by taking advantage of specific observer perceptions or other cognitive biases. Proper application of signalling system types to real-world cases will be sensitive to evolutionary timescale and ecological context. Telling them apart may be difficult without significant epistemic investment.

For our purposes, there are indeed important (and potentially tractable) differences to investigate between different signalling regimes, especially when seen through the lens of evolutionary trajectories and human-relevant timescales. As stated, hard-to-fake signals can have efficacy costs imposed by physical necessity, but their costliness has no adaptive function in itself – meaning that evolution will do what it can to minimise them. In a costly-fakeable signalling system though, (differential) signal costs are incentivised under certain conditions. CREDs mechanisms in comparison have costly display behaviour propagating (leveraging cognitive biases) but see no incentive for *differential* costs and benefits. So, while these explanations for costly displays might be difficult to tease apart empirically (especially without historical or longitudinal data), there are real differences between them in terms of evolutionary dynamics and empirical prediction.

In terms of evolutionary narrative and evolvability we can also see specific nuances and differences. For example, the pathway from a cue to an index is a relatively easy one: all that it would take is the evolution of some sort of reveal/conceal behaviour. Once a cue is capable of being recognised and responded to, simply putting oneself on display (in the right circumstances, under the right conditions) is in effect an index signal. Once this happens, evolutionary logic should favour such signals: “once an index is established in a population, it may no longer be an option not to use it. A stag that did not roar would probably be treated as of low quality” (Maynard-Smith and Harper 2003, 47). Robert Frank calls this the ‘full

disclosure principle' (Frank 1988), and it is often seen as giving index signals a more robust basis than fakeable signals, since all senders would have incentive to reveal their underlying type and no 'code' need be coordinated upon. Such signals can therefore easily originate as ways to exaggerate or make more salient an impressive trait, and the advantages of doing so will then feed into the usual mechanisms (natural selection, sexual selection, etc.) for evolution toward an ecological optimum[50].

A related, arguably more significant difference is that repertoire of unfakeable signals will be limited, as there must be a constraint or natural 'honesty-guaranteeing' mechanism connecting each trait and signal pairing. Selection can probe the search space for these connections/constraints, but they must exist and be accessible. For example, a stag's roar can signify its size and (to a certain extent) vigour, but not other, more subtle traits that would make it a formidable opponent or worthy ally/partner (e.g. ruthlessness, experience, or cunning). A capacity for loyalty/commitment, the trait we are interested in humans for RST purposes, would seem on the face of it to be relatively opaque as well, requiring a specific mechanism to make it diaphanous (as in Robert Frank's theory of hard-to-fake emotions, which will be explored in chapter 6 in application to ritual signalling). In any case, the need for 'hard' constraints as a fulcrum for hard-to-fake signalling is effectively the requirement for an entrenched, hard-coded infrastructure, the features of which define the scope for signalling options.

Fakeable signals, being more arbitrary, are more open to taking arbitrary forms, and more open to signifying for subtle or more hidden underlying traits (as long as the appropriate incentives exist). This is not to say that subtle and interesting unfakeable signals are not possible, for example it has been argued that bright, colourful ornamentations in some animals (especially males) might serve as a signal of resistance to parasites (William D. Hamilton and Zuk 1982) as they are hard to maintain under a heavy parasite load. But there are other subtle influences of signal expected dynamics. For example, a genuinely arbitrary, open-ended starting point for fakeable signals poses its own problems for evolvability – populations with more available arbitrary signal content will be slower (and less reliable) in how they narrow down their options

[50] Though see (J. P. Bruner 2015a) which demonstrates that the full disclosure principle is not as robust as might be thought, and under certain conditions (where high types are more prevalent than low types, and signals have minor costs) an uninformative equilibria is evolutionarily stable. This adds to the point that verbal descriptions and informal intuitions often turn out to be inaccurate with regard to expectations of when signalling must or should occur.

to a single mapping from trait to signal/action, even if well-incentivised to do so. In contrast, the need for underlying, independently constrained correlations means that index signals will depend to some degree on the evolvability of their enabling constraint mechanisms.

Returning to the explanatory desiderata from the previous chapter then, the two systems predict qualitative differences in signal characteristics. Fakeable signals could in principle 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. Unfakeable systems in contrast would be more 'athletics-like': ways of directly revealing/demonstrating underlying traits via a more constrained (and perhaps more generatively entrenched) diaphanous connection between signifier and signal, implying greater human commonality and less complexity and flexibility. If nothing else, this makes the space of differential cost-benefit signalling especially promising for application to the RST context.

4.4.2. Summary and context

This all means that we should avoid characterising signalling (including in the RST) via costly signalling or index signalling in isolation. While signalling theories are unified by their *need* for an honest signalling model, the available honest signalling models are quite disjunctive. They differ in their dynamics, limitations, and explanatory/predictive potential, albeit in some cases only subtly and in ways that are difficult to empirically discern. These multiple modelling options and their complexities all mean that serious attempts to connect real-world, signal-like behaviours to signalling models will be demanding.

To take one final biological example, consider the 'push up' display of the *Anolis cristatellus* lizard (Bennett and Huey 1990). Upon spotting a predator, the lizard becomes stationary and moves its body up and down by flexing and extending its legs, in what looks like a 'push-up' display. This seems like a clear case of signalling, but the signal form (and what is being signalled) is not obvious. It might index basic information about the lizard's condition: it being strong and healthy, so not a complete walkover. But it might instead work as a costly fakeable signal, signifying (for example) having plenty of energy in reserve for effective fight or flight – a tired lizard could put on the same display, just at greater risk in the case of being pursued anyway (having wasted some of its limited reserves). Why exactly a signal is being sent, i.e. which fact is being signalling for in what way, is not immediately obvious, but the exact

evolutionary story behind it will depend on these fine details of cost-benefit and signalling mechanism.

In short, the job of this chapter has been to open up questions rather than to offer answers. Later chapters will more systematically approach those answers. Models of signalling have been verbally described, but without the level of formal detail that would help us pin down exactly how they work and how they compare to one another. Qualitative differences between them have been similarly described, but without specific application to religious signalling or review of how they have been applied. I have also outlined several requirements and recommendations for such applications, but in the absence of an overall checklist or ‘recipe’ for doing so. Filling in the details here is the task of the next two chapters. By their conclusion, we will be in a position to divide the remaining open questions into empirical questions to be settled elsewhere, and questions that can be addressed by way of formal modelling – some of which will then be pursued.

5. Signalling Models for Religious Signalling

In the previous chapter I introduced several modelling distinctions with regard to signalling, but without formally defining any models themselves. The plan in this chapter is to fill in some of the formal details and use the greater precision this allows to take a deeper dive into the potential uses and pitfalls of these models for signalling theories of religion. Along the way, I will be formally defining the models which were introduced in the last chapter, and which will be applied to RST and used and modified in those that follow. These include the David Lewis signalling game of common interest, as well as models of differential cost and differential benefit signalling with partial conflict of interest.

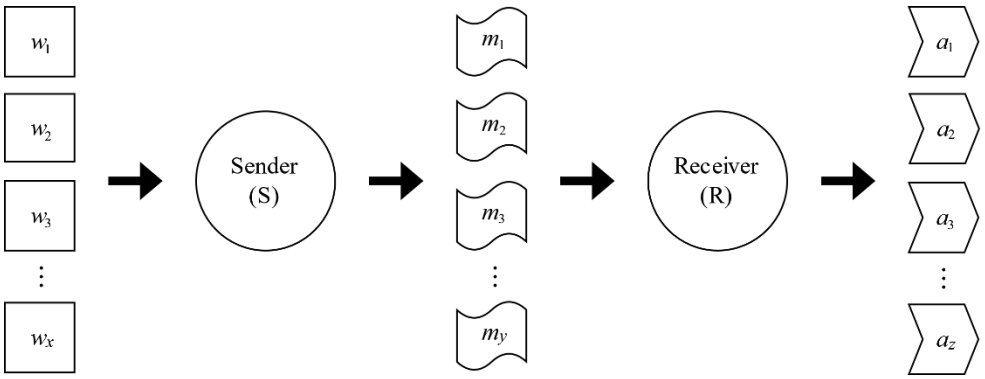
5.1. The sender-receiver framework

The sender-receiver framework just outlined is broadly attributable to philosopher David Lewis's account of conventional, intentional meaning in his doctoral dissertation and subsequent book (Lewis 1969). Precursors with something like a sender-receiver framework include Shannon's seminal early work on information theory (Shannon 1948) and in Peirce's theory of signs. Lewis's general use of the game-theoretic framework since been developed using evolutionary game theory with an evolutionary, naturalistic interpretation by Bryan Skyrms (Skyrms 2010) and other authors. The modelling methodology used in later chapters will follow that of the Skyrms school[51].

In the sender-receiver framework, senders and receivers are players in a dyadic game, where the payoffs for both players depend to some degree on some underlying, pre-determined state of the world (such as a trait of the sender), which had a set of possible states $\{w_1, w_2, w_3, \ldots w_x\}$. The sender moves first out of a set of possible signal or 'message' moves $\{m_1, m_2, m_3, \ldots m_y\}$, after observing which state of the world is actual. The receiver then has a range of possible action moves $\{a_1, a_1, a_3, \ldots a_z\}$, which they select between after observing the sender's signal/message, but with no direct information about the underlying state of the world. This is illustrated in figure 5-1.

[51] For example (Huttegger et al. 2010; Huttegger and Zollman 2013), and see (Godfrey-Smith 2011) for a quick historical context of the sender-receiver framework.

Figure 5-1: Diagram of the sender-receiver framework for signalling



Given what they can observe and what they can do, senders and receivers each have a set of *strategies* available to them, which include conditional strategies such as "m_2 if w_1, otherwise m_1" or "a_1 if m_2, otherwise a_2", and unconditional strategies such as "m_2 always" or "a_1 always". There are $i = y^x$ of these sender strategies $\{s_1, s_2, s_3, \ldots s_i\}$ and $j = z^y$ receiver strategies $\{r_1, r_2, r_3, \ldots r_j\}$, which can be thought of as the set of possible functions from the agent's input stimulus and their output response. However, these are just the so-called 'pure' strategies, 'mixed' strategies are also possible where the agent randomises between two or more of their available (pure) strategies according to some probability distribution.

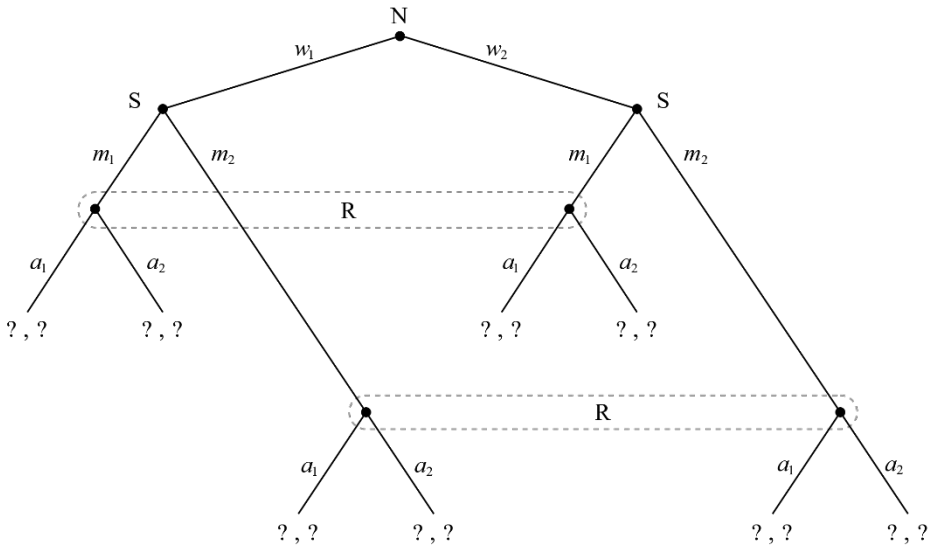
Signalling systems are possible where both sender and receiver use conditional strategies to some degree, but only when a conditional strategy pairing is mutually incentivised, i.e. those strategies are in an equilibrium[52]. In *separating* equilibria, the strategies are such that each *w* has been assigned a unique *m*, which triggers a unique optimal response *a*, such that neither sender nor receiver would do better by changing to a different strategy. Other equilibria types are also possible include *pooling* equilibria, where either senders or receivers act unconditionally, so that no communication takes place, and *partial pooling* equilibria, where conditional signalling occurs but world states do not all receive unique signals. Which strategy pairings constitute equilibria depend on the relative expected payoffs for each agent, and whether it would pay for either of them to change strategy (holding the other agent's strategy fixed). Typically, the sender's payoff is a function of their own move, the receiver's move, and the underlying fact, while the receiver's payoff is determined only by their own move and the

[52] Strictly speaking, the appropriate solution/equilibrium concept for signalling games is 'Perfect Bayesian Nash equilibrium'. One recommended introduction to game theory (at this level of detail) is (Gibbons 1992).

underlying fact (i.e. a particular signal might be costly to the sender, but not the receiver). It is always in the interests of the receiver to discern the underlying fact based on the sender's action, but it may or may not be in the interests of the sender to transmit any information about the underlying fact to the receiver.

The simplest models in this framework are '2x2x2' models, i.e. where $x = y = z = 2$ and there are two possible underlying states, two possible actions for the sender, and two possible actions for the receiver. The general template for such games is illustrated in figure 5-2.

Figure 5-2: Template for 2x2x2 signalling games, with payoffs unspecified



All games with this skeleton begin with a move by nature, or 'the universe', N (treated as a player for this purpose) which determines which of the possible states w_1 or w_2 is actual. The sender S then performs actions m_1 or m_2, and the receiver R performs actions a_1 or a_2. Crucially, when the receiver acts, they cannot distinguish between being on branches w_1m_1 or w_2m_1, and between w_1m_2 and w_2m_2 – only the sender's action, m_1 or m_2 is visible (the dotted bars connecting the relevant nodes are the 'information sets' of R). The eight possible outcomes of the game correspond to the eight combinations of *w*, *m* and *a*, and the payoffs assigned to sender and receiver at the terminal nodes (question marks in this diagram) determine which they should prefer to arrive at (the universe, as always, is indifferent).

Filling in this payoff structure with actual payoffs (numbers, or functions of other variables) defines the exact of game we are talking about. The basic 2x2x2 form can also obviously be

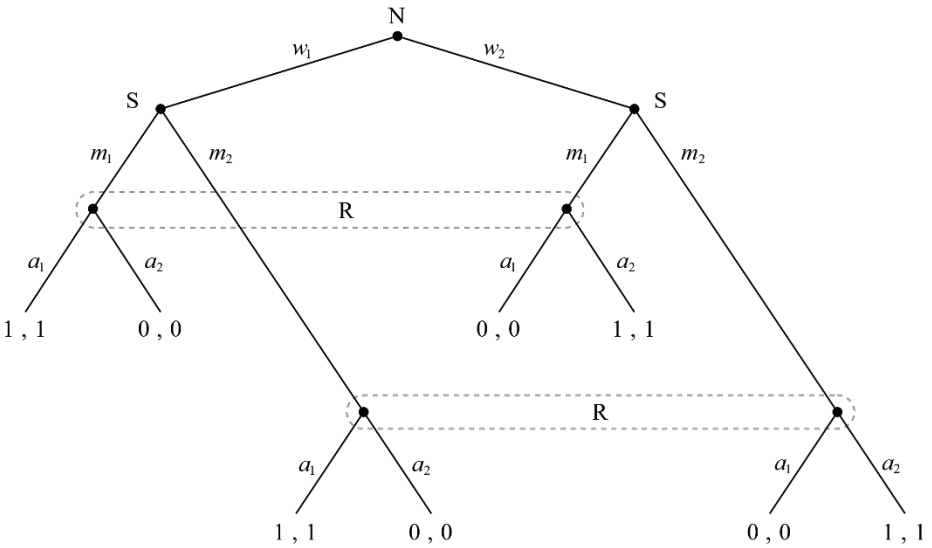
added to, for example with more states of the world and signalling options, and with probabilities assigned to non-player moves (e.g. probability of world-state, or probability of signal failure). These payoffs and other details then determine the communicative equilibria (if any) that each game allows, with the evolutionary significance of those equilibria to be investigated by computational or analytical methods.

5.1.1. Costless signals and the David Lewis signalling game

Lewis's account of intentional meaning modelled signalling as a common interest signalling game, with communicative conventions as the expected solutions to coordination problems. David Lewis signalling games are attractive in their intuitive simplicity and clear outcomes. They are coordination games of common interest between world-observing senders and action-making receivers using costless signals; in contrast to games where interests may differ and where costly signals are invoked.

The 2x2x2 David Lewis signalling game is shown in figure 5-3, which fills out the figure 5-2 game structure with specified payoffs. In this game, the sender observes that the world is in one of two possible states and broadcasts one of two possible signals which are observed by the receiver, who performs one of two possible actions. If the acts match the state of the world (i.e. a_1 if w_1 and a_2 if w_2) then both players receive a greater payoff than otherwise – there is no conflict of interest.

Figure 5-3: The 2x2x2 David Lewis signalling game with common interests



There are two possible combined strategy pairings between senders and receivers which get both players what they want:

1. Sender plays m_1 if w_1, & m_2 otherwise; Receiver plays a_1 if m_1, & a_2 otherwise.
2. Sender plays m_2 if w_1, & m_1 otherwise; Receiver plays a_2 if m_1, & a_1 otherwise.

For example, if the world is in state w_2 and the players use the strategies in line 2, then the sender will conditionalize on this to send signal m_1 and the receiver will conditionalize on the signal to perform action a_2. The action matches the state of the world, so both players receive the 'success' or coordination payoff of 1, and this is true of both strategy pairings whatever the state of the world – the only difference between the two strategy pairings is which states of the world w_1 and w_2 the signals m_1 and m_2 are correlated with. This provides the incentive the players need to stick with their strategies. There is no way that either agent can improve their payoff to unilaterally change their strategy, so these two pairings are Nash equilibria – separating equilibria, since they involve conditional strategies from both players. In Lewis's parlance (which I have already co-opted), these are *signalling systems*, and – most importantly – they are they only Nash equilibria of the game. In other words, these are the only evolutionarily stable solutions to the game; and they occur when senders condition otherwise arbitrary signalling behaviour on the state of the world, and receivers act on those signals to secure their mutual payoff. The implication is that populations which are capable of evolving toward a signalling system will *inevitably do so*, given enough time immersed in strategic situations which suitably approximate this payoff structure[53].

This result means that the David Lewis signalling game can serve as an ideal model (in a baseline modelling sense), from which other signalling models deviate in various ways. Alterations which stay inside the David Lewis game family of models include increasing the size of the game in different ways (i.e. beyond 2x2x2), and introducing imperfect information, mixed strategies, and so-forth. What is distinctive of David Lewis signalling games are the common interests of senders and receivers, and so other forms of signalling game can be seen as deviations away from correlated payoffs and common interest.

---

[53] Again, the evolutionary modelling methodology which demonstrates this will be introduced in chapter 7. This statement also assumes an unbiased world: i.e. that nature's moves w_1 and w_2 are equally likely.

5.1.2. Differential payoff signalling with partial conflict of interest

Having common interest baked into their payoff structure, the David Lewis signalling games do not suitably approximate the kind of strategic situation we were interested in; they are not directly relevant to the context of the cooperation problem. For partial conflict of interest, consider a signalling game which has correlated payoffs on the w_1 branch like the David Lewis game (i.e. either 1,1 or 0,0), but where payoffs on the w_2 branch are anti-correlated (i.e. either 1,0 or 0,1).; this fits a situation where there are 'high type' senders whose interests match those of the receiver, and low types whose interests are in conflict. Examples in the literature are numerous: species with profligate males of different quality and choosy females (females only want to mate with high quality males, but both high and low quality males want to mate with females), predators and prey (predators want to chase easy-to-catch prey and leave the superior specimens alone, whereas neither type of prey want to be chased), and human applicants with varying qualifications and commitment (the choosers want to admit only the best quality applicant, whereas all applicants want to be chosen). The high types on branch w_1 might start sending a signal to indicate their type, and receivers might start responding to it, but as soon as they do it will be in the interest of the low types on w_2 to send that same signal to 'lie' to the receivers. We can see that there are no separating equilibria in this signalling game, and similar dynamics play out here as in the prisoner's dilemma[54].

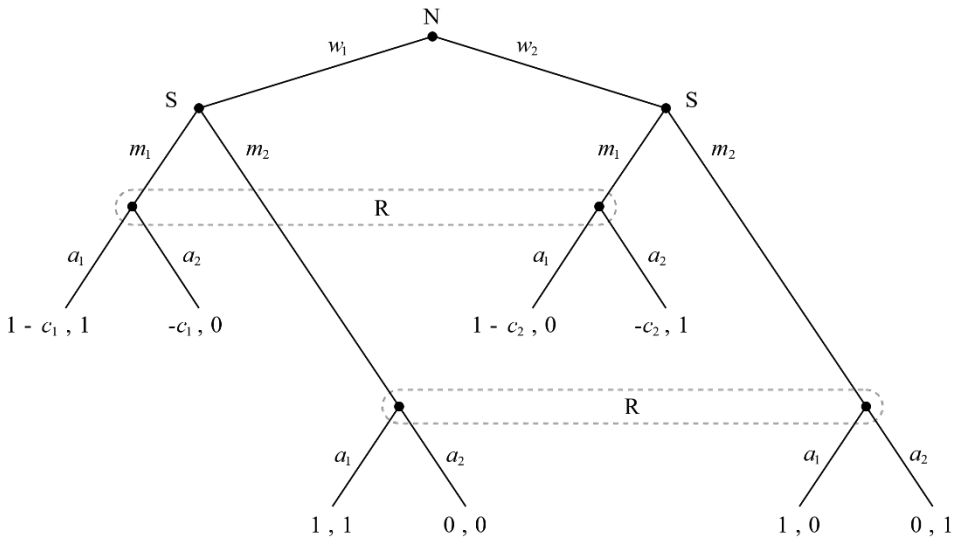
To get a game with a separating equilibrium though we need only add differential costs to the signal m_1. Consider the stotting example. The two possible underling states are that gazelle is either a fast runner likely to outrun the predator (high type, w_1) or more likely to be caught if pursued (low type, w_2). It has the option to leap and run (m_1), or simply run (m_2), while the predator has the choice to refrain from pursuing them (a_1, treating them as a high type), or to give chase (m_2, treating them as a low type). To leap is to send a costly signal of being a high type and the signal is kept honest by the differential cost, cheaper for high types (c_1) than for low (c_2).

The extended-form payoff structure which captures this is illustrated in figure 5-4. For simplicity, values of 1 and 0 are again assigned to the best and worst outcomes for the gazelle

[54] In fact, if we awarded nature's move to player one as well, it would be mathematically equivalent to a prisoner's dilemma were player one pre-commits to cooperating and defecting, but then gets to send a 'cheap talk' signal before player two makes their move.

(being left alone vs being chased) and the predator (getting the type of the sender right or getting it wrong). In this game the interests of sender and receiver are correlated in the case that the sender is a high type (w_1 branch), but anti-correlated for low types (w_2 branch), i.e. there is partial conflict of interest. On top of this, senders pay a signal cost on signalling branches (m_1), but this differs depending on the w-branch: the signalling cost for high types (c_1) different from that for low types (c_2). In the special case that $c_1 = c_2 = 0$, we have a conflict of interest game with no separating equilibria, but a separating equilibrium exists in the case that $c_2 > 1 > c_1$ (i.e. condition [4.2]).

Figure 5-4: 2x2x2 differential intrinsic-cost signalling game (partial conflict of interest)



Given the formulation of the game here, the strategies for this equilibrium are:

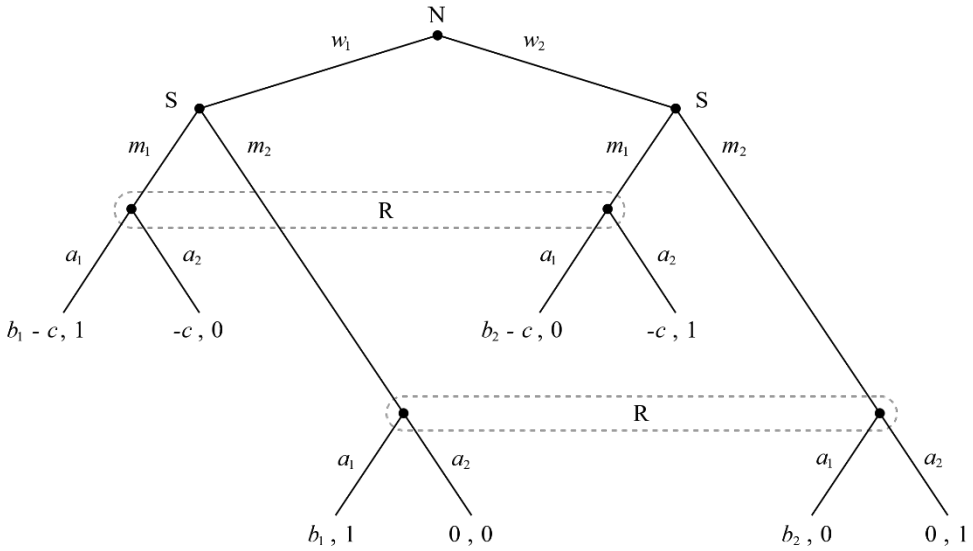
1. Sender plays m_1 if w_1, & m_2 otherwise; Receiver plays a_1 if m_1, & a_2 otherwise.

The key terminal nodes of this game are $w_1m_1a_1$, which must deliver a positive payoff for the sender (i.e. $1 - c_1 > 0$), and $w_2m_1a_1$, for which $1 - c_2$ must be negative. Under these conditions, neither sender nor receiver has incentive to deviate from their strategies. This simplified model is not the only way to model handicap signalling, but is used for example by (Zollman, Bergstrom, and Huttegger 2013), (S. M. Huttegger and Zollman 2013) and (J. P. Bruner, Brusse, and Kalkman 2017), and will be used here as well as a representative model for costly signalling.

Figure 5-5 instead shows a simple differential-benefit signalling game. In this signalling game, the cost for signalling (c) is now uniform and payoffs for the senders instead depend on two

new variables. Variable b_1, is the benefit that w_1-branch, high-type senders receive from being treated as high types (a_1), whereas b_2 is the benefit for low types being treated as high types. So high and low types pay the same signalling cost but receive different benefits for being treated as a high type. A separating equilibrium exists in the case that $b_1 - c > 0$ and $b_2 - c < 0$ (Zollman, Bergstrom, and Huttegger 2013), equivalent to condition [4.3].

Figure 5-5: 2x2x2 differential benefit signalling game (partial conflict of interest)



Despite the perhaps natural intuition that costs and benefits are just two sides of the same coin, these two games are not mathematically equivalent. For example, the sender payoffs in the $w_1m_1a_2$ and $w_2m_1a_2$ outcomes are different in the differential cost game (because the signal costs for high and low types are different), but the same in the differential benefit game (because being treated as a low type means no success benefit to mark out a difference). But it is easy to see that a similar function is performed in each game: differential payoffs between the two types of senders allow a parameter space for a separating equilibrium. And we can easily imagine merging 5-4 and 5-5 into a general differential payoff cost structure with both differential costs and differential benefits (not pictured). This would be similar to the payoff structure in 5-5, except with the c terms replaced by c_1 and c_2 on the w_1 and w_2 branches, as per 5-4. In this signalling game we can derive the general conditions for the separating equilibrium as per [4.1]. I.e.: $b_l - c_l < 0$, and $b_h - c_h > 0$, given our interpretation of w_1 and w_2 as sender being high type and low type. It is in this sense that differential cost and differential benefit signalling games can be seen as special cases of the more general model, where we either set $b_1 = b_2$ or $c_1 = c_2$.

Another more complex, ‘blended’ game which blurs differential cost and differential benefit is the Sir Philip Sydney game, introduced by Maynard Smith in (Maynard Smith 1991) and subsequently widely used as a model for costly or handicap signalling. In that game, the sender (‘signaller’) pays a cost c for signalling as a high type but the receiver (‘responder’) also pays a (fixed) cost d if they respond as if the sender were that type: it models transfer of resources from receiver to sender, with the benefit of the transfer for the sender depending on the sender’s type. In the usual interpretations of this game, ‘high type’ implies the sender is hungry, injured, or otherwise vulnerable in some sense such that receiving the resource is more valuable than for low type senders. Canonically (though somewhat confusingly) this is modelled negatively, with a being the harm suffered by w_1 high types if the transfer is not made, and b for w_2 low types, where $a > b$ – i.e. high types are ‘high’ in the sense that they are more in need of the resource transfer.

If this were all there was to the Sir Phillip Sydney game then there would be no separating equilibria available, as it would never pay for the responder to transfer in response to the signal. However, payoffs for the two agents are also cross weighted via a relatedness parameter r in the style of inclusive fitness. It will therefore pay for the sender to transfer the resource if interests are sufficiently linked, such that $-ra > c$, and there will be a separating equilibrium as long as d and b are low enough that this linkage does not either i) deter high types from drawing the resource away from the receiver and ii) incentivise unconditional transfer.

Because of the reliance on relatedness, this is not a signalling game that will be used further here. In part, it is a legacy of the original biological context of the handicap principle. But as a form of differential cost-benefit signalling, it also illustrates how broad that class of models can be, and how loosely terms like ‘costly signalling’ have been applied. The breadth of the general category might otherwise be missed here, with the focus on simpler models. Differential cost and differential benefit models should be recognised 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 signalling systems should approximate these instead of more blended forms satisfying condition [4.1].

From a formal modelling perspective, treating all these models as more or less fungible is reasonable: the models and their equilibria are similar enough to be considered all of a kind for abstract purposes. Loosely referring to them all as ‘costly signalling’ or ‘handicap signalling’

models is relatively harmless at this level of enquiry. However, when it comes to interpreting signalling models and applying them to putative real-world phenomena, there can be significant differences, many of which are not obvious.

5.2. More complex models & equilibria

One difference that can now be noted is that the evolutionarily stable equilibria of these models are more complex than the *separating* equilibria description given so far. The simple alternative to a separating equilibrium is a *pooling* equilibrium, where either the sender or receiver follows an unconditional strategy: always sending, always treating as low type etc. These equilibria are obviously non-communicative, but are common except in idealised cases, such as in the David Lewis signalling game with symmetric world-states. Also available in some games are *partial pooling* equilibria where communicative strategies are imperfect, and *hybrid* equilibria, where communicative strategies are mixed with unconditional strategies (J. P. Bruner, Brusse, and Kalkman 2017; Kane and Zollman 2015; Zollman, Bergstrom, and Huttegger 2013).

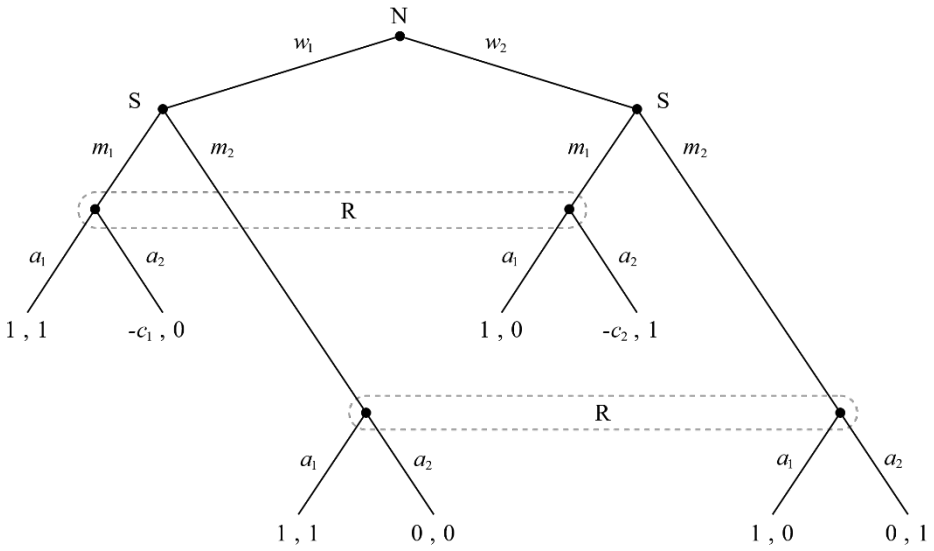
As mentioned earlier, partial pooling equilibria occur in games larger than 2x2x2, where there are opportunities for strategies which fail to map the relevant inputs and outputs on a one-to-one basis (e.g. a sender re-using the same signal for two world states, in the three world-state game). Equilibria composed of these imperfect communicative strategies are stable when either sender or receiver unilaterally modified their strategies would decrease their fitness and can be thought of as 'local maxima' on an uneven fitness landscape. Partial pooling equilibria also exist in unfakeable signalling systems often produced by the presence of efficacy costs (Jovanovic 1982), with these costs sometimes even quashing the evolution of signalling altogether, despite the so-called 'full disclosure principle' (J. P. Bruner 2015a). These results demonstrate the frequent inadequacy of even educated intuitions regarding what signalling theory predicts.

One particularly interesting case of this sort of non-obvious complexity and nuance involves hybrid equilibria and 'vulnerability signalling' games. I will argue in this section that some supposedly canonical examples of handicap signalling are better described instead by a vulnerability signalling model – which only allows for hybrid equilibria rather than separating. This confusion occurs (so I argue) because of imprecision and equivocation with respect to costs and how they are paid; which therefore connects these formal models with the earlier discussion of cost and the currency problem.

5.2.1. Vulnerability signalling and hybrid equilibria

The simplest vulnerability cost signalling game is similar to the differential cost signalling games of figure 5-4, but where signal costs are absent on a_1 branches. I.e. the sender pays a differential signalling cost depending on type/world state, but only if the receiver treats them as a low type[55]. This game is shown in figure 5-6. One biological example here that fits the vulnerability signalling model is the predator–prey signal (or 'predator approach' signal) sent by the guppy *poecilia reticulate*. The guppy, upon spotting a predator, moves toward it in a signal which indicates it is not the prey that the predator should be pursuing (similar to stotting). The signal has no cost though, unless the predator calls its bluff (via a_2).

Figure 5-6: 2x2x2 differential vulnerability-cost signalling game (partial conflict of interest)



This small difference results in significantly different equilibria, operating over significantly different parameter spaces (J. P. Bruner, Brusse, and Kalkman 2017). The signalling game has no separating equilibrium, but does have an evolutionarily significant, partially honest *hybrid* equilibrium, where senders and receivers mix together pure conditional and unconditional strategies, in proportions which depend on the proportion of high and low types in the sender population.

[55] In Searcy and Nowicki's terms, these are 'receiver-dependent' signal costs, as opposed to 'receiver-independent' (Searcy and Nowicki 2005), and so do not occur on every branch of the game where the signal itself is present.

Hybrid equilibria also exist in differential cost and differential benefit signalling games alongside separating and pooling equilibria, as described by (Zollman, Bergstrom, and Huttegger 2013) in a paper which first argued for their significance in the recent literature (see also (Zollman 2013)). The hybrids in these games exist only in a limited range of probabilities for types and strategy mixing. Let x be the prior probability of senders being a high type (i.e. probability of w_1). Senders employ an unconditional "signal m_1, regardless of type" strategy in a fixed proportion of their interactions = α (i.e. α is between 0 and 1), otherwise employing conditional strategy "signal m_1 if high/w_1, otherwise m_2". On the receiver side, β represents the probability of strategy "if m_1 then a_1 (treat as high type), otherwise a_2", with the alternative being "always a_2". Zollman et al.'s hybrid equilibrium exists when three conditions are met:

$$\alpha = \frac{x}{1-x}$$

$$c_2 = \beta$$

$$c_1 \leq c_2$$

The first condition requires that x be less than 0.5 (as probability α would exceed 1 otherwise – note that this is not generalisable beyond the 2x2x2 game). Overall, the second and third conditions correspond to a parameter space where the cost to a low type signaller (c_2) is greater than that to a high type (c_1), as with the pure strategy separating equilibrium, but where c_2 is below 1 (since β is a probability, i.e. between 0 and 1)[56].

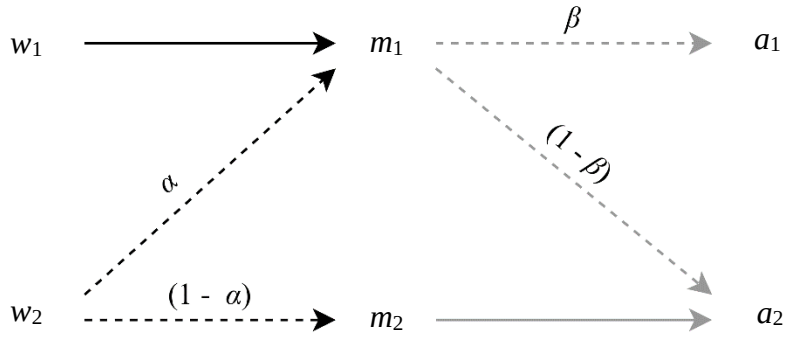
Figure 5-7 reproduces a sender-receiver diagram from (Zollman, Bergstrom, and Huttegger 2013), which illustrates these mixing strategies of senders and receivers at the hybrid equilibrium. At this equilibrium the strategies are:

1. Sender plays m_1 if w_1, & mixes between m_1 at with probability α and m_2 at (1- α) otherwise;

   Receiver mixes between a_1 at probability β and a_2 at probability (1 - β) if m_1, & plays a_2 otherwise.

[56] The existence of the hybrid is also constrained to populations where the signaller being a w_1 high type is constrained by the maximum value of α, meaning that the probability/frequency of high types must be low (0.5 when α is 1 and lower otherwise).

Figure 5-7: **Sender-receiver diagram for mixed strategies in the hybrid equilibrium.** Senders are in black, receivers are in grey, solid lines are conditional messages/actions, and dotted lines are probabilistic/mixed actions according to probabilities α *and* β.



In the vulnerability signalling game there is no classic signalling equilibrium, as "always signal" is a strictly dominant strategy in the case of the conditional response strategy "if m_1 then a_1 (treat as high type), otherwise a_2". In other words, there is no incentive for honest signalling, even if the receiver is primed to respond favourably to it: all types are incentivised to signal as though they are high types. In this respect then, this game more closely resembles costless signalling than costly signalling. However, again, this is only the case when considering pure strategies, and once we look at mixed strategies again, a hybrid equilibrium analogous to that found by Zollman et al can be seen.

As before, the hybrid has the signaller mix "always signal high" with probability α, and "signal high if actually high" with probability (1-α), while the receiver mixes "treat high if signalled high" at β with "always treat as low" at (1-β). The payoffs that the receiver sees are the same in both games, so, as with the standard differential cost game hybrid, one condition of the vulnerability cost hybrid equilibrium is also:

$$\alpha = \frac{x}{1 - x}$$

Again, another condition is that $c_1 \leq c_2$. But the different sender payoffs mean that equilibrium conditions for the sender are not the same. In this case, the additional cost condition is that is:

$$\beta = \frac{c_2}{1 + c_2}$$

Or alternatively:

$$c_2 = \frac{\beta}{1 - \beta}$$

This more complicated relationship between c_2 and β has interesting implications. In the differential cost game, the hybrid equilibrium is 'capped' by the standard separating equilibrium. As c_2 approaches 1, so must the likelihood of action a_1 in the face of m_1: converging to the information-preserving separating equilibrium when m_1 is always met by a_1. In vulnerability game hybrid equilibria, the crucial low-type vulnerability signal cost c_2 is approximately equal to β when β is very low, but the relationship is not linear: c_2 reaches the unit benefit payoff of 1 at β=0.5, and is nine times that at β=0.9. As β tends to 1, c_2 tends to infinity. And because there is no upper limit for the hybrid, this also applies to c_1.

In interpretation, this means the vulnerability hybrid has similar properties as the standard differential cost hybrid in the low signal cost range: it could potentially drive the incremental scaling up of signal cost on low types. But it can also potentially explain the stability of truly drastic potential signalling costs. As long as there is a sufficiently high chance of being taken at your word if you signal that you are a high type, it would be rational to risk signal costs (in the case that you are not) which dwarf the benefits of being treated as a high type. Importantly, this applies for both high-type and low-type signal costs: all that matters for the availability of the hybrid is that the chances of being treated as a low type are very low, and that a low type would suffer a greater cost than a high type when that does happen.

5.2.2. Costs, expenditures, and the relevance of the vulnerability cost model

One way that the formal details here are significant is with respect to the earlier misgivings about costly signalling in the religious context. In the simple separating equilibrium, the signal cost for low types must exceed the benefit of being treated as a high type[57]. This was rightly flagged by (Henrich 2009) as potentially problematic: it implied that signal costs for low types must be very high, even at the very start of a co-evolutionary relationship between sender and receiver strategies. But for the hybrid equilibrium, a signal cost for low type senders can start generating some sort of communicative result from a very low starting point. It is therefore much easier to imagine an evolutionary narrative where small costs start being introduced, without having to be matched by initially small communicative benefits. An obvious communicative benefit might therefore be approached in a hybrid equilibrium-based signalling system, with a gradual scaling up of costs via cultural 'mutations', until (perhaps) the parameter

[57] For the simple, 2x2x2 model at least; other more complicated signalling games might be able to recover some sort of signalling at the low-cost parameter ranges, but they will not be considered here.

space for a separating equilibrium is entered. Indeed, this is plausible at least from a modelling perspective: (J. P. Bruner, Brusse, and Kalkman 2017) also demonstrate that the evolutionary stability of the hybrid improves as c_2 increases, and (Kane and Zollman 2015) have also been shown hybrid equilibria to be potentially more evolutionarily significant than simpler communicative equilibria types.

These properties have the potential to explain bizarre religious costs, providing they are rare. This might not seem obvious at first sight, and indeed it will involve a more revisionary approach to the interpretation and application of signalling theory to human cases. This will therefore be considered in the following chapter. But again, there are cleaner biological examples that can be considered for illustration.

The example already used is of the guppy *poecilia reticulate* and its predator-approach behaviour. Interpreted as a signal (i.e. in the context of an appropriate response from the predator), it is a signal that says something like “I’m not scared, I can outswim you”. The guppy pays nothing for the approach itself, but if the predator disregards the signal and attempts to pursue anyway, the guppy has placed itself at considerable risk (which is greater if it is bluffing about being swift).

Another good example here is the *Anolis cristatellus* lizard and its push-up display, from the previous chapter. Recall that this is also a response signal to a predator (importantly, it’s common predator, a species of snake). Instead of fleeing, the lizard puts on this display, which should serve only to delay it and expend energy it could have used to get away. But if the snake tends to respond to the signal by looking elsewhere for a meal, we might recognise this as a rational gamble in the model of a vulnerability signal. The sticking point is that the lizard has actually paid what looks like an up-front signal cost: the energy expenditure the signal involves. As before though, there are a range of possibilities here. Seen through the lens of a hard-to-fake signalling model, it could be an efficacy cost: the necessary price for demonstrating its fitness and physical condition. As a strategic signal, the question would be how to interpret the energy expenditure and compare it to the other costs and benefits it would need to be commensurate with. Following (Grose 2011), and considerations in (J. P. Bruner, Brusse, and Kalkman 2017) about the difference between costs and expenditure, I argue that the common currency problem bites deeply here; as the question of which model fits the case turns on whether energy expenditure is a cost commensurate with the average fitness costs of conflict.

Consider again the more famous example of a stotting gazelles. There are a number of proposed explanations for this behaviour (including many non-signalling explanations) with mixed evidence in their favour (Caro 1986a; 1986b). But there is at least some plausibility to the idea that it is an honest signal of the ability to evade predator pursuit, which predators recognise as such (FitzGibbon and Fanshawe 1988). And though (Maynard-Smith and Harper 2003) are non-committal, this is a favourite proposed case of handicap signalling, used as illustration of the handicap principle for example in (Zahavi and Zahavi 1997), and in numerous discussions of the costly signalling theory of religion[58].

As might be expected, there is also an index signalling interpretation available for stotting. If we simply look at whether the Gazelle stots at all then this certainly looks like a fakeable signal. But stotting is a performance that can (plausibly) admit of degrees. How well each gazelle performs, if leaping as hard as they can, will provide a reasonably direct indication of their physical condition and dexterity – i.e. an index signal for physical quality. Simply leaping up from among the grass (and surrounding animals) in rigid position also gives the predator a much better look at them, making visual cues of their physical condition more salient. So, might stotting instead be an index signal, with efficacy costs, and with signal honesty constrained by basic biomechanics? Which interpretation is best will be an empirical question, depending on which features of the performance are more closely attended to, and how fitness is impacted.

But assume for argument's sake that predators do not attend to the fine details of the signal, and only recognise if it is performed or not. It is more likely in that case that it is a fakeable signal, and the costs of stotting are strategic. How good a fit is the differential cost handicap signalling model of figure 5-4 as a model of stotting? I argue that it is not a good fit at all, on grounds of incommensurate costs: there is no reasonable interpretation of the signal cost that supports the payoff structure of 5-4. The objection here pivots on what the assigned payoffs for success and failure (1 and 0) represent. For the sender gazelle, a 0 means that they are being chased. This number represents a collection of risks rather than certain death (death is not certain, not even for a low type), but the real-world risks are significant and dire. One mistake and they're a goner, and even if they do escape, they may have been driven away from the safety of the herd, and so-forth. Even a reasonably confident high type will not be indifferent

[58] For example, in (McNamara 2006; Northover and Cohen 2017; Bulbulia et al. 2013; Bulbulia 2004b; Bulbulia and Sosis 2011; Sosis and Alcorta 2003; Hall and Gonzales 2016; Bulbulia 2004a; C. Alcorta 2017).

between being chased and not. Absent signalling costs then, we have a payoff scale running between 0 and 1 where 1 is life going on as normal and 0 is non-trivial significant risk of death. This is the scale which signal costings must be calibrated against: it is a *fitness* scale, and this is what makes plausible the evolutionary significance of the alleged signalling game.

What kind of cost does the stotting signal add? The most obvious claim about the costliness of stotting is that it involves an extra expenditure of energy – it is metabolically costly. However, in and of itself, the *energy* expended by the act of leaping is insignificant compared to the fitness-risk scale. Gazelles don't avoid leaping like as if it were the risk of death itself, but this is what one is implying by intrinsically assigning it a significant, non-zero cost on this scale[59] in order to shoe-horn the case into the 5-4 model.

As with the *Anolis* example, the gazelle's extra energy expenditure isn't occurring in a context-free situation, as the key difference it makes is in the gazelle's ability to evade pursuit. It also makes more of a difference in this respect for low types than for high types, thereby making plausible the differential cost of this signal. Indeed, this is explicitly the rationale given in traditional uses of this example: "Only a gazelle certain of its ability to outrun a predator dares squander its strength in this way" (Zahavi and Zahavi 1997, 6). There are also other plausible costs and risks of stotting which are intrinsically fitness-negligible, such as the loss of split-second manoeuvrability for evasiveness while in the air, the risk of stumbling on landing, but all of these (except perhaps intrinsic risk of injury) become non-negligible *only in the context of being pursued*. In other words, if the signal cost is interpreted as a fitness cost, the sender only pays the signal cost if the receiver treats it as a low type.

But this breaks the figure 5-4 handicap model; because the costs just identified are only costs on branches of the game where the receiver responds in a particular way: we should only assign a significant cost to the signal at nodes which include the receiver actually pursuing the sender, i.e. down a_2 branches.

In other words, if it is a fakeable signal then it is a vulnerability signalling game that is being played, not a classical handicap signalling game as generally assumed. In idealisation then (i.e.

---

[59] One comparison here that might be drawn is Richard Dawkins' discussion of his famous Life-Dinner principle (Dawkins and Krebs 1979). While we expect both predator and prey to adapt to each other, Dawkins claims the prey species will experience a greater evolutionary pressure than the predator species, since failure for the predator means going hungry for an extra day, while failing for the prey means death. Similarly, the fitness impact of

looking at 2x2x2 games) we should be looking at figure 5-7 instead of figure 5-4. And this makes a difference: there is now no separating equilibrium, and the hybrid equilibrium has a significantly different parameter space and evolutionary dynamics. In interpretation, the mixing strategy predicts that the behaviour of individual agents (for example, receiver response to a costly signal) will be highly unpredictable, compared to the more robust conditionalization of receivers in the separating equilibrium state, with behavioural probabilities only appearing in statistical aggregate. Finally (though this wasn't apparent in the static analysis), the hybrid equilibrium point is only dynamically stable: in evolutionary games (see chapters 7 and 8) the probabilities will oscillate in response to each other over time. This all means that even messy field data might be compatible with there being hybrid-form signalling present.

It should be stressed that mathematical implications like these, derived from idealised 2x2x2 signalling games, cannot be straightforwardly projected onto real populations. The main point of this section was rather an *in-principle* illustration that i) small differences with respect to modelling choice can generate widely-divergent predictions, and ii) the rationales for such modelling choices turn on fine details of interpretation, especially with regard to how relative costs and benefits are estimated. Details matter.

5.3. Summary: how details can help

So, the good news for RST is that there are many more modelling options available for signalling mechanisms than just 'costly signalling', and more models mean a greater diversity of target systems that can be potentially be modelled as evolved signalling systems. The bad news is that these considerations also suggest great complexity with respect to application, potentially demanding great attention to interpretive detail. And we should remain cognisant that this discussion is all at the level of mechanisms rather than big-picture theory. So, the plethora of signalling mechanisms in no way guarantees that they will feature significantly in the true evolutionary story of religion and society, nor that they would be its only heroes if so. Signalling 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 that might include close conceptual cousins (like CREDs), and other number of mechanisms from cultural evolution or the cognitive science of religion. Indeed, with complexity comes operationalizability and empirical tractability concerns. Interpretation issues like those just discussed these will be multiplied in complex human societies where there are many traits of

potential interest. This is especially applicable to complex rituals and participation procedures: different features or religious tropes can have distinct signalling functions, with different dynamics and evolutionary histories entail various trade-offs.

The more recent biological literature on signalling in ecology provides some sort of hope and perhaps a model to look to, in that signals are used as part of a multi-trait, multi-modal *causal complex* of signalling systems (Hebets et al. 2016; Hebets and Papaj 2005). In this literature, separate signalling structures (for example, different parts of insect legs) are interpreted as having different signalling roles which nevertheless work in concert in specific contexts (for example in mate selection, or predator signalling). Depending on the signal forms responsible, there may also be complex trade-offs to consider. One trade-off already discussed example is that differential cost-benefit and hard-to-fake signalling-systems operating on the same signal could put contradictory pressures on the evolution of signal cost. The animal communication literature also suggests properties of these overall complexes, such as having signals with redundant, overlapping, or pluripotent functions to manage overall trade-offs between robustness, efficiency and evolvability (Hebets et al. 2016).

Of course, the need for trade-offs in modelling is nothing new. Richard Levins famously highlighted the basic trade-off for population models, between precision, generality, and realism: we can have two of these in one model (if we're lucky), but not all three at the same time (Levins 1966; 1968). Models can sacrifice generality for precision and realism (type I), sacrifice realism for generality and precision (type II), or sacrifice precision for generality and realism (type III). This is likely true for modelling religion and its social dynamics as well. But distinct modelling strategies can be complementary. For example, detailed study of the costs and benefits associated with specific, localised religious ritual traditions (with deliberate attention to the currency problem) could potentially permit tests that to validate or falsify some quite detailed signalling models, and/or discriminate between them. This would be a type I model, with the goal perhaps being an in-principle validation of the signalling mechanism in one narrow context. Abstract formal models like the ones in this dissertation are arguably more like type II models – they offer some general, proof-of-concept distinctions and/or predictions but are too highly idealised to be straight-forwardly applied to real religious target systems, but (at least with respect to signal form) there is at least some attendance to precision. Then there are imprecise, general studies of broad ethnographic phenomena, searching for a general signature of causal phenomena at the population level, such as we saw in chapter two with (J.

Watts, Greenhill, et al. 2015) and other 'big-picture' studies with respect to the Big Gods theory. Another type III-style method could be conducted at the local level, looking at correlations between participation religious participation and long-term outcomes.

Each of these approaches can do complementary explanatory work. In the context of this dissertation and my earlier characterisation of it, the 'detail' project of chapter 3-8 can be given a Levins-style gloss as being focused on generality, and (formal) precision. The ongoing argument is that details matter, and so narrowing down on how signalling models are constructed and applied can open up potential new approaches (potentially of both type I and type II) that can shed light on the explanatory potential of signalling theory and perhaps (in application) use it to make some novel, testable predictions.

The question then arises: how much detail need RST really get into? I suggest that this depends on how precise an empirical science it aspires to be, and that 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 seems questionable that detailed ethnographic analysis of specific human religious traditions as signalling complexes could approach even this level of detail. But in principle there is no reason why some degree of progress could not be made, perhaps along the lines of the level of precision seen in recent work on social networks and religious signalling in village-based religious communities (Power 2017). The level of detail that formal approaches to signalling theory allow might outstrip its scientific testability in many contexts, but it may still help inform more targeted empirical approaches, and at least *some* of the nuances should be taken into theoretical account, especially if moving beyond simple verbal models.

6. Signalling theory applied

Thus far, I have argued i) for a core schematic characterisation of RST that is neutral between honest signalling models, and ii) that the choice of honest signalling models can make a significant difference in what a fully specified RST theory might end up saying. In this sixth chapter I apply the prior theoretical and formal lessons more specifically to the case of religious signalling. In the first section I look at general applications and interpretations of signalling to human societies. The second section argues that distinctions are being missed in some of the existing theoretical literature; distinctions which are important in pinning down RSTs and making them empirically tractable. I conclude in the third section by proposing some templates for more fully specifying RSTs, in the sense of interpreting signalling models such that they might be mapped with suitable precision onto real-world phenomena.

6.1. Human applications in general

In this dissertation I have largely avoided delving into the ethnographic richness of religious practice for examples evocative or suggestive of signalling. In part, this is because specific examples can be *too* suggestive. As discussed earlier, the chaotic, noisy Shia Tatbir/Talwar Zani public ritual of cutting one's own scalp with a sword until blood flows freely (an annual gift to xenophobic tabloids throughout the West), is a visually arresting and confronting example of how pain and sacrifice can be enthusiastically and proudly put on display to roars of approval. As Dimitris Xygalatas and co-authors have documented, self-mutilation and piercing with hooks and skewers takes on calm, deliberate, and sometimes artistic expression in the Kavadi, during the Hindu Thaipusam festival on island of Mauritius (Xygalatas et al. 2013). Religious ceremonies to initiate young people or consecrate marriages lasting for days are not uncommon throughout the world, with participants' displays of commitment (and endurance) carried out under intense public scrutiny. But far less extreme forms of devotional ritual and sacrifice are even more common. Alarming proportions of meagre personal resources are often handed over to religious authorities; restrictions on movement, dress, association and other freedoms are often intricate and jealously monitored; and large numbers of people gather weekly or daily in coordinated, very public expenditures of time praying or re-absorbing familiar diatribes.

Many behaviours look like signals being sent and received, some can be grouped by similarity, and many are intuitively costly in some sense, but there is no overall pattern with respect to form and detail. Focusing on certain areas of human religious behaviour can be suggestive of

this or that signal form (or other mechanism), but when looking at the ethnography of religious ritual (or even just dwelling on anecdotal observations), it becomes clear that the diversity of religious practice does not cleanly support any particular hypothesis over the other. But the diversity of signalling mechanisms suggests a general method: look at examples case-by-case to draw what lessons we can, without trying to shoe-horn them all into a single model. Signalling mechanisms are disjunctive in the abstract, the question is how disjunctive they might be in human reality.

6.1.1. Costs and interpretations, again

We can begin with a simple case that can help recap and further contextualise the warnings from previous chapters: a display of wealth in the form of a religious donation. Such an action is obviously costly to the agent, and such donations are generally observed and acknowledged, by religious authorities if not directly by one's peers. It might therefore be a costly signal of social/religious commitment, rewarded and incentivised by improved social standing. Alternatively, a display of wealth might just as well directly signal for *wealth*, i.e. social dominance or social standing – it is after all valuable to know who the local elites are, and for the elites make themselves known as such, and for those lower down the rung to mark out their own place on the ladder. This latter kind of signalling does not lend itself to the sort of signal-cooperation co-evolutionary story that we are interested in though, as elite status is famously distinct from prosocial commitment (Piff et al. 2012). As before, non-signalling explanations (CREDs, pyramid schemes, etc.) are of course possible as well.

In another example, consider rhythmic dancing or chanting rituals common in many traditional societies, which last for many hours and are a considerable test of endurance. If we were looking solely for a costly, arbitrary ritual (calibrating this with the assumption that the only trait of interest was prosocial tendency of some kind) then this might look exactly like what we expect to see: a costly ritual that makes no sense unless in the light of separating the team players from the flaky or the lazy. However, tests of endurance like this are also (surprise, surprise) pretty good tests of *endurance*: and endurance is a quality which is generally valuable in forager or other traditional economies: running down prey, roaming long distances, not whinging or slowing down the group during joint tasks, etc.

Unlike raw strength or aggression though, and more like commitment, physical endurance isn't a quality you can easily demonstrate on the spot – it is an intrinsically diachronic aptitude. As

such, you need a more drawn out means of demonstration and (preferably) a captive audience while doing it. Long mass rituals therefore make a certain amount of sense for testing/signalling for either quality. Endurance rituals are in-principle compatible with either a fakeable or an unfakeable signalling hypotheses, though one empirical test here might be to see if endurance rituals were disproportionally present in societies where endurance labour is economically vital. And, perhaps they serve both functions: hard-to-fake signals of endurance and costly signals of being a team player.

But here too, rituals requiring feats of endurance; the bearing of pain and so-forth are also compatible with a variety of non-signalling explanations which might complement or rival signalling mechanisms, depending on the details. One potential difference in prediction here is downstream outcome. Signalling-based explanations have it that the ritual in question is some sort of performance to be tested or assessed, with participants then filtered, sorted, or graded via standardised norms of performance. If there are no consequences for 'failing', then a signalling explanation is not supported – the signal must be incentivised by receiver-side responsiveness. It is fair to ask then about the behaviour on the receiver's side, i.e. how much sorting actually goes on, subsequent to the real-world rituals we want to apply the theory to.

For example, some of the most personally demanding religious rituals are initiation rites, some of which seem to have quite low refusal and fail rates, with participants simply forced to endure them (Morinis 1985). This suggests several possibilities. There might be an element of indoctrination to these rituals, and their primary purpose might be not to sort for the underlying qualities, but instead inculcate or reinforce them. It might be that observers remember the quality of ritual performance and discriminate accordingly on a longer-term basis; or perhaps in these traditions there was greater discrimination in the past and the observed ritual is a vestige of that (indeed, one of the results in chapter 8 will support this idea). Or perhaps everyone invests only just enough to pass. But the point is that the signalling hypotheses make more predictions than just about the causal history of a ritual, the proximate mechanism (including some kind of sorting behaviour) matters.

An alternative, broadly Durkheimian mechanism is 'social technology', which would build commitment rather than filtering for it. I.e. extreme ceremonies (especially those involving multiple participants) generating common experiences of hardship that in turn create greater commonality and familiarity between community-members, directly reinforcing in-group

cooperation and cohesion (Whitehouse 1995; 2018b; Atran and Henrich 2010; Fischer and Xygalatas 2014). Two closely related versions of this are what we might call the 'training' or 'boosting' hypotheses: both involving the idea that extreme rituals (and preparations for them) have the effect of altering the participants in way that makes them more useful to the community. Such a function might be just a short-lived exploitation of psychological traits evolved in another context (i.e. 'boosting'), or developmentally and/or culturally cumulative ('training'). Either way, the religious ritual would be introducing a socially useful phenotype modification that can figure in an explanation of that ritual (perhaps via cultural group selection or some other selection or learning mechanism). In a more bleak interpretation, the submission to such treatment (carefully controlled and stage-managed to incrementally build up over a developmental lifetime) might function to set a psychological precedent of sorts; desensitising participants to personal indignity and individualistic concerns, and raising the bar for what constitutes a normal level of submission.

Telling the difference between such hypotheses (especially multi-functional customs are a possibility) will generally be problematic unless ethnographic data is specifically gathered to tease them apart.

To take a concrete example, consider the ritual ordeal that teenage boys belonging to Kenya's Kalenjin people traditionally submit themselves to, as described in (Warner 2013)[60]:

> Elly Kipgogei, 19, remembers going through the ceremony at age 15.
> First, he says, he had to crawl mostly naked through a tunnel of African stinging nettles. Then he was beaten on the bony part of the ankle, then his knuckles were squeezed together, and then the formic acid from the stinging nettle was wiped onto his genitals.
> But all that was just warm-up; early one morning he was circumcised, with a sharp stick.
> During this whole process — the crawling, the beatings and the cutting — Kipgogei was obliged to be absolutely stoical, unflinching.
> He could not make a sound.
> Indeed, in some versions of this ceremony, mud is caked on the face and then the mud is allowed to dry. If a crack appears in the mud — your cheek may twitch, your forehead may crinkle — you get labeled a *kebitet* — a coward — and stigmatized by the whole community…

[60] It should be noted that this is a journalistic source, rather than peer-reviewed ethnography. But nothing much hangs on this, for current (purely illustrative) purposes.

> … After Kipgogei was circumcised, he wasn't allowed to go home. He was taken to a hut on the outskirts of the village to heal from the operation and he was told, whenever you leave this hut, you are not allowed to walk.
> "So you're supposed to run and it's very fast. So you're running very swift, having the pain," he said.

The price of failing to stoically endure this ordeal is great enough that young boys will practice and prepare for it, for example by burning themselves with hot coals. Stoic behaviour in this context looks like an indicator trait for both endurance and dedication, with the stigmatisation of the insufficiently stoic indeed suggests that it represents some sort of index signal being sent and tested for, on a pass-fail basis. It could also be a strategic signal of dedication (of a variety to be laid out later in this chapter). But of course, this could also be training, boosting, or ritual subjugation. Or some combination.

In terms of broad comparisons though, there are differences between the mechanisms that might leave traces worth looking for. First, the presumed predictive difference between sorting and manipulation explanations would be whether *all* boys take part in the process or some are indeed discarded or demoted by their community. E.g. a fakeable signalling hypothesis would predict that the less stoic and committed individuals (low types) would tend to simply avoid the ritual and accept second-class status rather than submit to the ritual. The ritual would function to filter out and discard these boys. Anecdotally, it appears that avoidance/refusal indeed occurs in contemporary Kalenjin society: some boys flee the ritual and the societal pipeline is leaky to some extent. But there are possible confounding factors: e.g. how much of this is an artefact of the contemporary context? Post-colonial Kenya offers alternatives for these boys that were not traditionally available, so the pipeline might be leaky by design or due to environmental mismatch.

Second, deeper cultural investigation might yield clues, but we would have to be wary not to take them at face value. At least according to anecdotal interpretations from this report, the point of the ordeal is not just to test for stoicism but also to instil it: after being forced to push on through that level of hardship they are thereafter far more tolerant of it and are more willing to forbear pain in the pursuit of valued goals. Of course, this might be post-facto justification and part of the culture's evolutionary design rather than any deep insight into it. But in (Warner 2013), it is considered as a possible reason why the Kalenjin are almost solely responsible for Kenya's sporting dominance in the marathon, and other forms of endurance running. On this interpretation, the Kalenjin's rituals are a highly successful system of social sanctioning which

forces individuals to train and condition themselves to meet community expectations of endurance, which also makes them disproportionately effective in endurance athletics as a side benefit. The signalling/sorting interpretation could also explain this: rather than a highly successful training program for endurance, the rituals might instead be a highly successful *selective breeding program* for endurance. But this yields a third set of predictive differences, this time with respect to genetic vs cultural heredity and trait-dependence, which suggest programs of empirical study. Can Kalenjin-specific genetic correlates be identified? Could an outsider be raised as a Kalenjin and also develop these traits?

So while signalling and training explanations are *not* mutually exclusive, as their combined use in (Fischer and Xygalatas 2014)'s model shows, there is at least scope for getting empirical traction on whether one is dominant, and to what degree.

6.1.2. Mixed signals

There are no clear substantive conclusions to draw from the Kalenjin case (not from my brief glance at it at least), but it illustrates how complicated the analysis of the function of ritual can be, and how little can be said definitively in the absence of detailed empirical study designed to separate competing hypotheses. It also highlights the sort of assumptions which should be avoided, to avoid begging the question in favour of the costly (fakeable) signalling hypothesis. These deserve to be summarised and collated together with points made in earlier chapters.

First, it would be a mistake to assume that prosocial dispositions are the sole candidates for the traits that communities might be interested in testing for via ritual. Even assuming that participation *is* explained by a signalling system of social quality, it would be hard to tell if it were i) a differential cost-benefit fakeable 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 an entirely different desirable trait, for example that one has the fortitude and physical capacities to be a *useful* member of that group. Indeed, in any cooperative endeavour, commitment is only one socially desirable trait among many, and without other redeeming features is often not desirable at all. Anyone at all familiar with committee work, amateur sports teams, or community orchestras will be aware that the most committed are not necessarily the most useful and, below a certain level of aptitude, a high level of commitment can instead become a liability.

Second, signalling alternatives are *not* mutually exclusive: in some cases we might see both i) unfakeable signals of some valuable quality and ii) costly signals of being willing to pitch in with it. Following the above reasoning, it makes no sense to test for commitment alone. The ideal team member is not just temperamentally inclined to pitch in but also capable of doing so at a high standard, and the value of commitment is dependent on it being concomitant with 'hard' qualities like strength, skill, or talent.

Third, we should not assume that the presence of putative costs or hardships are in and of themselves evidence of costly *fakeable* signals, ala the handicap principle, or indeed of signalling at all. As argued in the previous chapters, we have to show that there is not just cost, but the *right type* of cost to be additive with respect to benefits on offer, when understood under ap appropriate signalling model, and at a level that hinges on underlying quality/type of the sender (depending on the model). Again, this is not a specific to the RST, with similar issues seen in applications of signalling across the human sciences (Grose 2011) and in biology (as argued in chapter 5) where theorists leap to signalling models too rapidly.

Fourth, there are multiple alternatives to signalling. We might assume that a costly display must be in some sense 'worth it' for senders in the long run, otherwise the ritual system would not be stable. But this is to assume that it is not just maladaptive, and neither a by-product nor the result of fitness-orthogonal mechanisms. And these are many and varied possibilities. Telling between them would take careful contextual examination, including what's happening on the receiver side – who they are, what they get out of it, and how might they be shaping the strategic environment. Without such information, a given cost, hardship, or expenditure might easily be:

a) underwriting the honesty of a fakeable signalling system (of some form),
b) the unavoidable price of generating or enabling an unfakeable signal (efficacy cost),
c) part of a program of training or directed psychological development, perhaps the result of cultural group selection where both the costly conformity and punishment that sustains it are individually maladaptive,
d) a wasteful, maladaptive by-product of cognitive biases with no benefit to anyone (CREDs and HADD explanations fall into this category),
e) a wasteful by-product of the sender's pre-evolved responses being manipulated for other purposes (like the wasp being manipulated by the orchid),

f) a transfer of resources from the participant to the community, either as a ritualised way to achieve economies of scale (e.g. paying into an investment society or 'paying forward' into communal activities like barn raising), or to non-adaptively (at the community level) enrich those already in a position of greater power,
g) an out-of-equilibrium phenomena, which will fade as signalling is gradually selected against,

And of course:

h) any combination of the above.

Finally, the importance of option h) should not be understated. None of a) to g) are mutually exclusive. It might hurt the *prima facie* case for the RST that there are multiple possible explanations for costly rituals (with a tricky empirical task to separate them), but it is simply not the case that the presence of (for example) a boosting or training effect means that the costs responsible for that are not also doing other work. Indeed, we might positively expect multi-functionality of ritual, because a multi-functional ritual system is likely to be more stable (via redundancy) than any similar system with a single function (ala the model of signalling complexes in ecological signalling). And this fits the target system. Religious rituals are not atomic and have many 'moving parts': stages, steps and features which can each be individually attended and responded to. Each part might have its own discrete cause (or causes plural). Again, this is to be expected: the history of evolutionary biology suggests that re-use and multiple use of complex evolved structures is more likely to be a norm than an exception. There is no reason to expect culturally evolved rituals and other cultural features to be any different in this respect.

I would argue though that the real lesson to draw from the plethora of possibilities and messy cases is not that the RST mechanisms are less likely to exist, but that expecting nice clean examples to support them (and only them) is naive. There is also positive reason to expect multi-modal signals which might advertise a variety of receiver-relative sender qualities (i.e. beyond just prosocial commitment). As argued, since both commitment and aptitude would be valuable to a community, there is no reason to expect that only one of these be their focus. But while commitment is a highly fakeable trait, many of the basic aptitudes which make it valuable in the first place are far less opaque. So while we might expect a mix of signals in the wild (if signalling is indeed a real phenomenon) we might also expect a correlation between form and

function, i.e. fakeable signals (kept honest by costs or other means) for commitment, and hard-to-fake displays that index and make salient the hard aptitudes which give commitment its value.

This is mirrored to some extent in the biological literature, where (Számadó 2011) writes: "The diversity of honest signalling and of the mechanisms that maintain it, producing cost-free, minimal-cost signals and handicaps at the equilibrium, will reveal itself in its full shape and beauty when researchers are primed to look for it[61]." In a detailed analysis of several previous studies into bird coloration (Weaver, Koch, and Hill 2017) state that, despite (William D. Hamilton and Zuk 1982), the field of animal colouration studies has largely ignored indexical signalling hypotheses and focused on the handicap principle. They argue that despite much scientific research on various handicap and index signal models (for example into the roles of testosterone and energy investments in chromatin) the jury is still out regarding whether handicap or index models fit the data best, and that (similar to Számadó) the two approaches are not mutually exclusive. In neither of these papers though do the authors explicitly develop the idea that the same signal token might simultaneously be both a handicap signal and an index – that is, an index of one trait but a handicap signal for something else – or be composed of a mixture of mechanisms which alternately realise handicap and indexical signalling functions.

If anything, long, elaborate religious ritual traditions with many moving parts seem like ideal candidates for housing signalling 'complexes' like this, bundling together multiple functions and realiser mechanisms. The trick would be empirically validating this expectation.

6.1.3. The science and interpretation of religious signalling

To round off this section then, I will briefly consider some of the empirically focused literature on religious signalling. If the thinking so far is reasonable, then possible cases of religious signalling should ideally be investigated with an eye to separating out multiple potential signalling mechanisms from non-signalling mechanisms and other confounds. But these sorts of distinctions are not the typical focus of such work. One recent summary statement is as follows:

---

[61] Számadó here is using costs to categorise signal types, with 'minimal-cost' signals being indexical signals that require expenditure to generate, and cost-free signals are indexical signals that have only negligible generation costs.

> "Signaling theorists hypothesize that religious systems evolve to facilitate within-group cooperation because religious costs enable partners to reliably predict the cooperation of others." (J. H. Shaver and Bulbulia 2016)

As we've seen, this doesn't narrow things down much with regards to signal form and modelling options. Rather, the focus tends to be on the applied side. For (Bulbulia and Sosis 2011) and (Bulbulia 2011), religious rituals are described as commitment devices, which need to be underpinned by 'hypnotic' convictions which bypass rational utility calculation. Likewise, religious signals and beliefs are seen as 'charismatic' (Bulbulia 2010) and "motivate a vision of reality in which it pays to cooperate" (Bulbulia and Sosis 2011). This is building commitment rather than (or as well as) signalling commitment. I.e. signalling theory in the literature is already syncretic to some degree, often appealing to functions for signal-like displays which go beyond the mechanistic schema at its core.

However, the central idea of signalling, especially focusing on (perceived) signalling costs, informs at least some ethnographic fieldwork studies. As discussed in chapter two, some empirical studies have offered evidence that religious participation and religious priming correlates with higher rates of cooperation and community longevity. Recently, a handful of important ethnographic field studies have also been published which bear on the role of rituals as signals and coordination devices, and the distinctions drawn so far in this chapter.

One study of a religious festival in the Mauritius (Xygalatas et al. 2013) compares the financial donations made by participants and observers involved in low-ordeal religious rituals (singing and collective prayer) and high ordeal rituals (involving "body piercing with multiple needles, hooks and skewers" over several hours). The authors reported that both participants in and observers of the high-ordeal rituals showed much higher levels of donation and nationalistic self-identification compared to low-ordeal participants. At face value these results appear to support a training or boosting hypothesis, with the study offering "the first natural demonstration that suffering predicts prosociality by capitalizing on intense, real-world stimuli that would be hard to manipulate in the laboratory" (Fischer and Xygalatas 2014, 1604).

The results here should be treated with cautious optimism. Being a real-world observational study, the participants in the two ritual types were not randomised into low or high ordeal ritual conditions, nor were their behaviours and attitudes before and after the rituals compared for evidence of change. However self-report of overall religiosity and religious history was controlled for, and the associations remained statistically significant. Interestingly, self-reports

of pain (rather than the specifics of the cutting and piercing that took place) also correlated significantly with donation behaviour, suggesting a direct experiential cause. So, while it is still possible that the causation underlying this correlation went the other way, with the attendees who were more zealous *on the day* gravitating to the high-ordeal ritual and/or the less zealous inclined avoided it, this potential confound (a mismatch between self-attribution and occurrent motivation) would take further study to validate. However, one last potential complication here is that participants and observers in the high-ordeal ritual tended to be *related* to one another. This suggests that the high-ordeal rituals could well have been deeply personal and familial experiences as well as religious ones. Relatedness also means that it is not clear how surprised we should be to see such a correlation in their behaviour.

In short, while the data here is interesting and (to a certain degree) compelling, it is not entirely clear what is going on. And it is certainly also compatible with a signalling hypothesis of some kind, especially as we don't know whether (or how) the observers' responses feed into outcomes for participants. In particular, there is no information about how observer perceptions of the participants change as a result of their performance, nor about any subsequent interactions between them.

A more recent and far more complex study (Power 2017) was expressly designed to test for such effects, in a mixed Hindu and Christian Tamil population in two small villages in Tamil Nadu, India. The ritual religious practices in these villages are rich and varied, occurring at annual festivals and throughout the year including symbolic processions, ritual vows involving abstinence, fasting & avoiding conflict, food sacrifices, extreme rituals involving fire and piercing (including being suspended by hooks), and, in the case of some of the Hindu participants (either in an official role or spontaneously in groups) becoming "possessed, their bodies contorting wildly, beyond their control and consciousness as a deity suddenly "comes" to them (*cami vantatu*)". What Power investigates is the relationship between the costliness of rituals and how participants in those rituals were perceived by the other villagers, both immediately after the relevant festivals and in the longer term.

Two aspects of this study are of interest for current purposes. One (more philosophical) point of interest is that signalling theory here is very much treated as an undifferentiated whole. Indeed, when introducing the theoretical motivations for the study, Power uses the term 'honest signalling', and 'signalling theory' rather than costly signalling theory. However costs were

also treated as a key determinant of the study design, in a way that invokes the handicap principle: “The signaling theory of religion places import on the differential costliness of the acts carried out” (Power 2017, 86). Different ritual types were ranked (statistically, according to villager responses to a card-based survey tool) according to costliness, and this ranking was used in the analysis[62].

The second is the findings: Power found that villagers’ perception of each other were indeed sensitive to and shaped by observed religious participation. More religious participation of a costlier nature correlated positively with several reputational qualities, especially being evaluated as ‘devout’ but also (in particular) being a hard worker, but also (to a lesser extent) other assessments such as strength, ritual knowledge, generosity, and good character. One character assessment which did not positively correlate was influence. Interestingly as well, regular worship was more effective than one-off displays: a costly display not backed up by ongoing, less costly demonstrations of devotion was relatively ineffective.

These results are novel and important, but also difficult to interpret. For our purposes though we can extract some relevant observations. First, we can divide up Power’s breakdown of reputational qualities into qualities which speak (somewhat) to the basic signalling models we’ve been considering. The fact that attributions of being strong are positively correlated with ritual participation suggest that at least some part of the signal being received was not concerned with prosocial traits: physical aptitudes too are on display, or at least being imputed. The most straightforwardly prosocial quality included was generosity, followed by ‘hard worker’ – these too correlated positively, ‘hard worker’ more so than generosity. ‘Religious devotion’ and ‘ritual knowledge’ might be associated with group loyalty, but both are also straightforwardly reasonable descriptive assessments to make about someone observed as taking part in many religious activities. As Power herself notes, the ‘religious devotion’ result is also in agreement with a CREDs interpretation. And, while a multiplicity of traits are seemingly being looked for, this says nothing about whether it is evolved signalling which is advertising those traits (as opposed to by-produced cues), nor which types of displays or signalling systems (if any) are driving the evolution of these institutions of ritual.

---

[62] though it is not clear, at least to me, how these costs are interpretable as differential between the committed and uncommitted.

For the moment then, the empirical evidence remains suggestive but not conclusive. But the methods being used and developed are promising. For example, Power's research also includes a remarkably thorough database of relationships and attitudes that the villagers hold with regard to each other, and how these change over time. And techniques like these hold great potential for highly precise, realistic models of religious signalling in this localised setting (i.e. Levins-style type I modelling), with potential to cross-validate results from less realistic, more general formal modelling.

6.2. Applications in the early religious signalling literature

In any case, some of the more abstract complexities of signalling models (and the possibilities they make available) have only been made salient with recent modelling work, so it would be understandable if it had not been appreciated, at least in the early RST literature. Indeed, according to one literature survey, the use of signalling 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), though this is greatly influenced by Robert Frank (Frank 1988) for his understanding of signalling.

6.2.1. Frank's hard-to-fake emotional displays, and religion

Irons draws heavily Robert Frank's 'Passions within Reason" (Frank 1988), which contains one of the most influential examples of an index-signalling model in human behaviour. Frank argues that emotional responses such as facial expressions, tone of voice and body language are evolved signalling mechanisms: hard-to-fake signals of motivational states and commitments to action. As cues, observers can pick up on these manifestations quite easily, but when they are strategically deployed by evolved behaviour and/or intentional display (posturing so as to make them salient), they are best seen as strategic signals. They are reliable because it takes considerable effort (and acting talent) to convincingly simulate anger, joy, and so-forth, especially when in the grip of a countervailing emotion. And they are strategic in the sense that they can make credible certain threats and promises, such as threats of violence, that would be less credible if merely issued verbally. Therefore, they provide the basis for reasonable inferences about the sender's behaviour, which in turn make a cautious observer's

responses more likely to be suitably correlated, and so (given certain incentives) this is the basis of a signalling system that might be reinforced by evolution or learning.

Applying this explanatory mechanism to intense, public displays indicative of normative commitment (and other moral emotions) is an obvious step, though of course these attempts will have to be assessed based on their merits[63]. As discussed in chapter 3, one solution to a commitment problem is to 'throw away the steering wheel', as it were – to visibly and unambiguously alter your own set of strategy options such that doubting your commitment to a given strategy is unreasonable. Flying into a righteous rage is a clear analogy to this, in analogous (e.g. 'chicken'-like) strategic situations – a credible loss of emotional control and rational calculation which indicates a low likelihood of backing down, or high likelihood of costly retaliation. Hard-to-fake signals of motivating emotions reveal that the sender is liable to act in certain ways, irrespective of what would otherwise be in their immediate, short-term interests, and so provide reliable signals of being committed to those actions.

In prisoner's dilemma-like strategic situations the best collective outcome comes from mutual cooperation (especially in iterated situations), but the best individual strategy at any given instant is to defect, so the commitment problem is to convince cooperation partners to cooperate in the face of this. Communities of *conditionally* cooperation-committed individuals, who recognise each other as such via transparent, unfakeable moral emotions, would be able to conditionalize their behaviour upon these signals and so avoid the mutually sub-optimal outcome as well as free-riding; thereby stabilizing cooperation and its benefits and so in turn incentivise the relevant moral-emotional traits.

However this sketch of a signalling system, a virtuous evolutionary feedback loop, is highly idealised. Emotions are messy. The obvious doubt here is the degree of unfakeability of moral emotions: intense emotions like rage might be difficult to fake, but more everyday moral emotional displays, such as displaying concern with the interests or problems of others are eminently fakeable (many friendships would not last long if they were not). Even keeping that to one side, while people seem to be very good at discerning emotional states in a one-to-one situation, and while different emotional signals appear to be strongly cross-cultural with a few

[63] Other than Frank's own work, the basic model has been invoked in a variety of ways (with various degrees of plausibility), in explanations of emotional traits such as grief (Winegard et al. 2014), aspects of romantic love and jealousy (Buss 2016), and aggression and vengeance in war (Boster, Yost, and Peeke 2003).

exceptions, discerning emotion is not the same as discerning the motivation behind it. There is not an *observationally* special category of moral emotions. Genuine, boiling outrage might be hard to fake, and its immediate target might be obvious, but the emotional background and triggers of such emotions – *why* we find the outrageous thing so upsetting – are typically less transparent. And unless the behavioural outcome of it is contained in the *immediate* future, it is unclear at best how clean the connection is between outrage and future behaviour. Frank's model works most plausibly for in-the-now interactions.

This might be exacerbated in a more formal, ritual, performative, or institutionalised setting which complicates one-to-one emotional 'connection', at least in the absence of entirely honest verbalisation of emotional outbursts. Is the village elder getting worked up by the alleged transgressions of my neighbour, or about some transgression of my own that I've unwittingly revealed in making those allegations, or just because of me bothering him about it? Is the emotional state of the penitent transgressor a case of genuine moral regret, or regret at being caught, or is it just shame or embarrassment at being observed at all? In the case of regret at least, both the reliability and usefulness of such judgements in the courtroom have been questioned (Bandes 2016), in the sense of being able to accurately discern regret, and in the sense of regret accurately predicting reduced recidivism. Emotional displays elicited and observed in the heat of the moment might well serve as clear signals of inner emotional states, but add layers of formal, social context and staging (such as inquisitorial institutions or scheduled, ritualised displays of commitment) and it is doubtful that emotional performances in these contexts are equally diaphanous.

So, there are four points of reasonable doubt here: i) are the relevant moral emotions sufficiently, reliably unfakeable, ii) are the relevant moral emotions sufficiently reliably discriminable from other emotional states, iii) are the relevant moral emotions sufficiently reliable indices of desirable traits at all, and iv) does any of this scale up to be useful in the appropriate group-settings? If we are looking for a signalling mechanism fine-grained enough for the task of driving the evolution of complex human cooperation, the information that emotional signals convey might well be so incomplete (in his context) as to make them a dubious candidate.

Importantly, the Frank family of mechanisms would also require a strong conditionalization faculty on the sender-side: it would require an emotional attachment and motivation toward

norms, i.e. rules for applying behaviours rather than just behaviours themselves, and most importantly rules for who they should perform those behaviours around. A passion (transparent or not) for unconditional altruism is not a trait we should expect to last very long in a heterogenous population: it must be a highly selective passion for being an altruist with/among other altruists (and presumably co-motivate altruistic punishment of non-altruists, once altruism becomes a norm). In total then, this is quite a lot of ducks for a blind evolutionary process to get all in a row.

This scepticism is in no way a knock-down argument[64], but it maps out the sort of case that would have to be made for a Frank-style mechanism (for our purposes). There is undoubtedly something true about the emotions-as-signals story, the question is: in what social worlds did it play an evolutionary role? Pre-human hominid social worlds for example are dominated by hierarchies built up from dyadic relationships. This setting (and moving into the early hominin lineage) seems like the natural place for unfakably emotional signalling to play a role, for example the emotions associated with friendships might be seen as positively assorting in the appropriate manner.

Likewise, small interaction groups where cooperation/accommodation is incentivised by high coordination dividends (large game hunting, collective defence, co-parenting, etc) make perfect sense for the evolution of i) norm-adherence, and ii) diaphanous emotional responses to preferred conventions and their violation. This is especially true for short-term, 'what are we doing this morning' time scales. All things being equal, it is better for me to assort with conspecifics who are 'on the same page' in that we have compatible approaches toward communal work. Displays of annoyance and anger with conspecifics who aren't going about the work 'properly' can serve as a reliable way to signal conventional expectations and individual senses of what proper coordination entails; allowing observers to better identify compatible social partners. And, in a well-assorted group, similar displays by the majority can send a reliable message that further violations of the convention or norm would risk withdrawal of accommodation or even costly retaliation of some kind.

But again, this illustrates the 'scaling up' problem – how far does an emotional 'connection' sustain itself in a one-to-many or many-to-one context? As I have suggested, a move from

[64] Nor is much of it particularly original, Frank's view has been around a while and quite thoroughly chewed over.

immediate, face-to-face, 'in the moment' emotional interaction to a ritualised, displaced setting for emotional displays, and/or the demand for more subtle and nuanced moral emotions to be expressed, make the reliability of those displays more suspect. In larger groups (i.e. outside the immediate interaction group where emotional pictures of conspecifics can be built up over prolonged observation), the evidential benefits of emotional displays will be more stochastic and less reliable.

In any case, applying Frank's theory to religious ritual is another obvious move, articulated (or at least appealed to) in various forms in the literature (Irons 1996; 2001; Sosis and Alcorta 2003). The general verbal model goes something like the following: given that a community values an individual's desire and commitment to take/maintain their place among them (as opposed to just taking what they can get), an evolutionary plausible trajectory for that community would be towards requiring ordeals that publicly elicit indicative emotional responses. Again, this would involve hard-to-fake, commitment-related emotions, and observer/community responses which were appropriately sensitive for mutual incentivisation of all the relevant strategies and behaviours.

6.2.2. Irons and the early development of RST

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 signalling 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, 293).

However, the terminology is ambiguous. In (Irons 1996) he characterises 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, 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 fakeable 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. 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 signalling 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 2012a):

> “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 ‘in the moment’ cognitive resources. What counts against the success of low types is the potential failure of the charade when their resources run out and quite high error rates even when the resources are being invested; 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 signalling theory in general (J. P. Bruner, Brusse, and Kalkman 2017; Kotiaho 2001), which has also been explored in the RST context (Murray and Moore 2009), and is part of the aforementioned 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

signalling proper (in the sense of the handicap principle) stages an infiltration 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, in part because the phrase "costly-to-fake" does not appear in (Zahavi and Zahavi 1997), but also because a move has now been made 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 signalling with hard-to-fake signalling. In a later paper Sosis also cites (Johnstone 1997) to say that: "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, 100). It is telling that Johnstone's discussion of signal cost here was only in the context of Zahavi and differential-cost handicap signalling, not hard-to-fake signalling. 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 behaviour is driven by a differential *perception* of costs by the sincere and the cynical, and perception-driven selection of signalling strategies is subsequently proposed as the driver for the evolution of religious signalling (Bulbulia and 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. Internalised beliefs and commitments might adjust *perceived* payoff and qualitative experience of participation, but objective fitness payoffs are less plausibly impacted. Informational signalling-like behaviour without material fitness differences[65] 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 core, adaptive RST framework this theory is something of a hybrid: allowing adaptive co-evolution of signalling and cooperation (via positive assortment) but with non-adaptive sender-side signalling strategies. It is still recognisable as a

[65] In this I am taking the idea of *merely* perceived costs differences seriously. It is possible for example that strong difference in preference carry with it difference in opportunity cost – the true believer would be at church anyway, even if their signal was not observed. But this is to re-describe the case instead of considering it as is.

variation on the general theme but arguably blurs the lines between RST and alternative, fitness-orthogonal, views such as CREDs theory. Again, it also forfeits one of the parsimony virtues of a pure signalling theory (operating entirely at the level of fitness and adaptive selection).

Sosis is clearly aware of the distinction between handicap and index signalling in (Sosis 2006), and by (Bulbulia and Sosis 2011) the picture defended appears to largely appeal 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 and Sosis 2011, 366). However, the nomenclature here is still idiosyncratic and 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 signalling model and incongruent with the biological/formal literature. But the later paper appears to opt for 'hard to fake' signalling, 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 and Sosis 2011, 365). However, shortly afterwards they also cite (Zahavi and Zahavi 1997) as a guide and emphasise 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 and Sosis 2011, 366–67).

The intention to not to treat these papers as a single authorial unit for pedantic finger-pointing with respect to terminological consistency, but rather to illustrate two conceptual points that are key to understanding what a successful modelling strategy here might look like. First, it is clear that Bulbulia and Sosis are interested in *honest* signalling in the broadly evolutionary sense, incentivised by and co-evolving with increased cooperative assortment. Despite emphasis on particular terms and concepts, they apparently have no axe to grind with respect to the two main branches of signalling. And this is entirely appropriate: any kind of signalling system will do for RST (to a first approximation). But at some point 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 fakeable 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 realiser mechanisms with different potential explanatory virtues and distinct evolutionary trajectories (especially regarding signal costs). They should not be conflated or treated as fungible.

Second, these papers illustrate two ways of responding to the differential objectivity challenge. One option is to drop objectivity in favour of 'hybrid'-RST handicap signalling with subjective cost differentials. The other is to drop differential costs in favour of non-handicap, diaphanous index signalling based on honest-emotions or other cognitive constraints. Again, there is no need to pick just one of these options, but these are also *not* the only options. I will conclude the chapter by outlining several potential templates for RST, based on cognitive constraints, differential costs, and differential benefits.

6.3. Mapping models to target systems

This brings us to the penultimate theoretical step: we need a plausible set of rules for how to apply signalling theory to religion. 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. The charge to avoid here is …… it being so vague and 'anything goes' that it resembles an unfalsifiable Panglossian program (Gould and Lewontin 1978) – a charge that has been levelled at the Zahavis (Pomiankowski and Iwasa 1998). Vaguely, verbally shoe-horning religious rituals into signalling models shouldn't be good enough. We need some sort idea of a procedure and set of standards for mapping signalling models to target systems.

6.3.1. Models and target systems

There are some theoretical resources to draw upon for this. Signalling models are just mathematical structures, and following Michael Weisberg's picture of scientific modelling (Weisberg 2013), explanatory scientific models should also include an *assignment* mapping of the structure's key features to features of a target system: the agents, the signal, the signified trait, the response, and the real-world mechanisms or institutions which correspond to game theoretic strategies and underpin their evolution. Taking the STR seriously therefore means not just appealing to signalling theory but also assigning the mathematical structures of signalling models to target systems in a plausible manner.

One of the simplest assignments would be a mapping for a public, mass religious ritual as follows:

Signal: participation in the ritual (whatever that entails),

Senders: the active ritual participants (plus those who decline, i.e. don't signal),

Receivers: observing conspecifics (including fellow participants),

Response: subsequent treatment (with respect to cooperative assortment),

Mechanism: behavioural norms and practices (encoding strategies), inherited vertically from parent to offspring, or horizontally via some form of success-sensitive cultural learning.

This assignment provides the minimum level detail needed for a tractable hypothesis, when paired with an interpreted signalling model: for example differential cost-benefit with some type of payoff structure (i.e. actual costs and benefits for the people in the sender & receiver roles), or index-signalling based on some tractable signalling constraint. More complex target systems and assignments are of course possible if some of the above features are varied (e.g. with senders and receivers being elites and followers rather than peers), and some of these combinations will correspond to existing STR variations from the literature. But in any case there needs to be some level of interpretation and assignment. If the formal and theoretical distinctions between signalling models and payoff structures are complex, their application to real-world cases in religious signalling adds another layer of complexity.

The overall recipe for a viable evolutionary model of signalling would therefore look something like this: a) the costs and benefits of signalling and outcome must be able to be combined into overall payoffs which b) makes sense to feed into some plausible update mechanism, c) over an adaptive timescale that makes sense for that model and its interpretation in the target system.

In the stotting case the (tacit) adaptive mechanism and timeframe was natural selection over millions of years on the savanna. The (high type) gazelles and their predators are hypothesised to have a biologically evolved a signalling system, driven and maintained by the fitness incentives of sending and responding to a signal, kept honest by its genuine, significant, and differential fitness costs for the senders. The problem in this case was the leap from energy expenditure being costly in one objective respect to it constituting a fitness cost in the strategic environment.

In the case of extreme religious rituals, the modelling situation is even more messy. Consider the possession ritual as described in (Power 2017). To model the devotees as senders in signalling game we must identify costs and benefits which combine into payoffs that are fit to feed into some evolutionary update mechanism; and do likewise with the observers whose changed perception of the sender collectively deliver the mutual social benefit. Presumably, the update mechanism is a cultural one, either laid down in development or via more ephemeral learning, reinforcement, or success-following. Fitness/prosperity benefits from social connection, and whatever costs are associated with the possession ritual must be made commensurate somehow in the context of this update mechanism, and this will require some sort of plausible/validated rationale. The theoretical gaps here are not be impossible to fill, but they would require work.

6.3.2. Payoff-update system

Most of the generic assignment elements listed above have been discussed already in previous chapters. The last element is the most complex: the evolutionary mechanism whereby sender and receiver strategies are encoded/embodied, such that they have some form of heredity but also sensitivity to selection-like processes that update them over time and allow them to evolve. This therefore links the previously discussed currency problem with an update requirement: there needs to be an overall payoff-update system that is coherent and plausible. This deserves to be unpacked in greater detail in the human cultural context.

As previously argued, what is crucial is that payoffs can be ranked and compared. This is what would allow strategies to have something like meaningful variability of relative fitness. Payoffs may have cardinality (with a numerical value on some scale or other, as I am using here), or simply have an ordinal ranking, but to generate meaningful results the preference orderings of both sender and receiver must each be complete and (to some extent) determinate[66]. This formal requirement has implications which need to be kept in mind when mapping costs and benefits of signals and displays (ritual or otherwise) to formal models of signalling. Game theory and its solution concepts are just mathematical tools, but you cannot link signalling and

[66] Agents can of course be *indifferent* between two or more outcomes, because their outcome payoffs at the nodes are (determinately) approximately the same. Indeterminacy or incommensurably between them on the hand would be a systemic problem.

the cooperation problem without them, and this means that there must be ‘playing fields’ which determine what’s at stake for the players.

Consider the following examples of payoff-update systems.

1. Pricing behaviour of two corporations in a duopoly. The stakes for the players are monetary, the motives are maximising profit, so we can draw predictions from pricing equilibria (assuming the players are free enough and clever enough to price accordingly) as the players adjust rationally to each other’s pricing moves.
2. Two human beings are able to subject each other to painful electrical shocks in a sadistic, 2-way variation on the Milgram experiments. The payoffs are comfort or discomfort, the motivation in minimising discomfort, and we can predict behaviour to the degree that the players actions are driven by hedonic self-interest (instead of anger or sadomasochism, for example).

Now imagine a version of (2) where the players can trade off money for electric shocks. The problem now isn’t just how to convert the ‘currencies’, it is also how to select the paradigm of rational strategy update. We could imagine (as economists do) a more abstract notion of utility evidenced by (or constituted by) the preferences revealed in action and decision, which monetary value and hedonic value can both serve as proxies for in cleaner cases like (1) and (2). But is there a principled way of combining both hedonic and monetary payoffs in a single utility scale, prior to running the experiment? Probably not, at least not in any simple or usefully general manner[67], which is probably why prudent experimental economists stick to a single measurable quantity (usually money) as a proxy for utility.

However, this is very similar to what seems to be going on in many verbal models of costly religious signalling. The costs of an extreme ritual are said to be pain and discomfort, whereas the benefits are such things as improved standing in the community and improved prospects for cooperation opportunities, and not much thought is given to how and why signalling strategies might be being composed and modified over time. Comfort, money, respect, and social standing are all ‘good’, in some sense, but more is needed. It’s very plausible to imagine

[67] We can imagine an experiment designed to reveal what the preference trade-off functions between comfort and money might look like. What are the marginal costs of pain avoidance? Is there a level of agony that no amount of money will compensate? Ethics committees alone make this epistemically inaccessible.

that financial burdens, the respect of your peers, and your standing in society might all have an impact on eventual reproductive success: they might imply indirect costs or benefits to fitness. However, they are indeed quite indirect, and there will be no simple, linear mapping between any of these into fitness. Furthermore, how does the pain inflicted by an extreme ritual, experienced a few times a year over a few hours or days, impact on fitness? Probably not a great deal at all, especially if the norms surrounding rituals encourage the participation of supporters to take care of the main participant, as they generally do for initiation ceremonies.

Perhaps we should instead see these costs and benefits as feeding into a utility maximisation or decision theoretic framework: we avoid pain and seek social and material benefit, and so at that level of abstraction we can indeed assign some sort of value. This must be the case to some extent: people act on a variety of motivations without suffering agential incoherence. But this has two problems. The first is as already mentioned: costs and benefits are being passed through a highly subjective filter before behaviour is generated, so differing proxy measures require conversion (somehow) before they can be useful in modelling that behaviour. E.g. pain is aversive, but is it aversive in the same way as going into debt is aversive? No, and it is reasonable to think that the cognitive systems which underpin pain avoidance work differently from those relating to debt avoidance (the modern literature on behavioural economics also shows that framing effects can make a huge difference in terms of which cost is acted on in any particular occasion). Second, if costs and benefits are translated into utilities, it becomes harder to link them to anything like a paradigmatic evolutionary, population-level feedback mechanism. Of course, this is only a problem with respect to a dependence on the biological signalling paradigm. As briefly mentioned in chapter four, signalling theory in economics takes its own form with analogous (but often distinct) modelling options, and there may well be tools there that can be co-opted instead.

Perhaps this is the direction that religious signalling theory should go, but that is not a question for this dissertation. I will simply assume that (at least some of) the lessons of more biologically attuned signalling models will be general enough to carry over. In any case, there are other adaptive paradigms available. As suggested in chapter 1, behavioural traits can also be transmitted both horizontally and vertically by mechanisms such as social learning and ‘follow success’ rules; and there are some plausible ways that these might approximate the effects of selective mechanisms (cultural fitness, Heyes-style cognitive gadgets, etc). Some sort of payoff-update system along these lines might be a viable option as well, though I will not

further elaborate on this either. Of course, there is also natural selection of traits transmitted via genic inheritance, though this may be more plausible for broader 'reactive' causal components of signalling strategies, such as temperament and basic social and aesthetic capacities or biases that can facilitate ritual displays and their reception.

Whatever the payoff-update system might be for a putative signalling system or signalling mechanism, we needn't be pedantic in insisting on a complete, independently validated story. But a coherent one is required. There will also be peculiarities to take account of; each possibility will imply its own scales of what constitutes success or failure, and therefore the features of the situation which correspond to costs and benefits. One other thing to bear in mind here (regarding the potential multiplicity of mechanisms) is that different evolutionary update mechanisms (natural selection, cultural processes, etc) will have corresponding differences in evolutionary timeframe. The consequences of this will be investigated further in chapters 7 and 8.

6.4. Modelling templates for religious signalling

Though the discussion thus far has been largely devoted to the complexities that a serious RST would have to grapple with, there are also positive recommendations to be made. But all the various decision points (which model, which traits, which costs/benefits etc) are not independent, and the search space for reasonable applications of signalling theory to religious signalling can be narrowed somewhat. Before concluding with a breakdown of the remaining open questions then, one set of these recommendations I will offer are some tentative 'templates' for modelling religious signalling. The idea here is to practice what thus far has only been preached; and try to bundle together some of the various options into plausible, potentially useful 'templates' for application.

6.4.1. Cognitive constraint templates (hard to fake)

Beginning with hard-to-fake signals (in the sense of index or constraint signals), we have already seen (via Sterelny) one interpretation where dishonest commitment signals become hard to fake: where lies either elicit too much cognitive dissonance or weave too tangled a web for the talents and efforts of the liar to reliably overcome. This will lean on the Frankian notion discussed earlier: that emotional/affective responses providing a hard-to-fake, honest window onto the true commitments of religious participants.

Consider the following reading. Intense or demanding religious rituals are culturally evolved ways of eliciting commitment-indexing affective responses: either hard-to-fake displays of enthusiasm and engagement from the genuinely committed, or hard-to-avoid giveaways of discomfort and dissonance from those only taking part for instrumental reasons. Taking part in these rituals might incur fitness-relevant costs, but they need not, e.g. experiencing pain during an extreme religious ritual is 'costly' in some intuitive sense of the word but might not translate into an evolutionarily relevant cost. Considering the evolutionary pressures on signal costs in such a model, one might therefore expect the evolution of this mode of religious signalling to optimise away from fitness-impacting demands, in favour of discomfort, pageantry, and awe.

This Frankian template for cognitive constraint-based religious signalling ticks several boxes. It includes a clear generative link between prosocial commitment and reliable signals. It 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 can be a dead giveaway). Extreme, painful rituals would also fit this template, given the apparent cognitive penetration of religious devotion into the experience of such rituals (Jegindø et al. 2013).

But it also has limitations. Revealing cognitive constraints by placing senders in a specific constructed situation is not obviously applicable outside the ritual context. As an index signalling model it does not predict seemingly arbitrary demands with fitness-significant costs; such as direct tithing or sacrifice of resources, or restrictions on dress, diet, economic activity, and sexual practices. And it does not explicitly predict signal complexity or cross-cultural diversity, as it is the emotional responses of the participants that constitute the signal, not the specific tropes and components of the rituals themselves. And any scepticism about the reliability, fakeability, or stability of moral emotional displays would have an impact here, see e.g. (Bandes 2016; Deem and Ramsey 2016). And in the abstract, unfakeable-signalling interpretations need exogenous constraints or connections to leverage, bringing with them both synchronic robustness requirements and the need for a further evolutionary back-story (as with cognitive biases in non-adaptive, non-signalling explanations).

Such considerations are not knock-down objections, and for signalling interpretations of religious ritual this is perhaps the most appropriate template to use, but they highlight gaps which differential cost-benefit mechanisms might instead fill.

6.4.2. Differential cost templates: punishment and bridge-burning

With respect to differential-cost fakeable signalling, we already have one option on the table: Sosis's hybrid model where the evolution of signalling strategies is driven by subjective perceptions of cost (rather than by objective cost-benefit), and the signalling (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 in response to Bulbulia), signals might map to persistent adherence to codes of dress or practice, perhaps where the absence of lapses or exceptions functions as the signal, rather than any one-off adherence event. In this vein, signalling strategies involving costly signals also need not be strictly 'pay as you go'. A comprehensive survey of all the possibilities will not be attempted, however, there are potential examples worth delving into with respect to temporally extended signalling.

The first comes with punishment of defection, which some argue should be folded into the notion of signal cost (Fraser 2011; Murray and Moore 2009). Suppose that acceptance into a community after signalling commitment now places you at the mercy of that community's retribution, should you defect from your commitment. In terms of cost-benefit analysis, such a signal of commitment followed by defection would be equivalent to signalling 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 modelling sender-receiver reactions, the order of associated cost and benefit is not: they only matter for purposes of determining overall payoffs. This means that the whole temporally extended scenario can be modelled as a single step in an evolutionary signalling game in which both high and low types pay the same (if anything) at the time the commitment signal is sent, but signalling low types also reliably pay an extra cost of being caught out.

Importantly, this is a signalling game with differential costs which meets Henrich's differential objectivity challenge and avoids the need for Sosis's retreat to subjective differential costs. The evolutionary proviso is that the overall behaviour of the community and the individual agent must be governed by strategies which are updated or reinforced in a way which approximates an adaptive response with respect to overall objective payoffs. If so then those strategies can co-evolve, and a stable signalling system emerge.

A related template is ‘bridge burning’. Consider the case of group-specific ritual scarification, tattoos, and other permanent markers. One powerful way of promoting group cohesion is parochial altruism (Bowles and 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 has made a powerful statement of commitment, because they have ‘burnt their bridges’ and dramatically increased their own cost of defection (Sterelny 2012a). If we again ignore the temporal ordering of cost and benefit and only look at the overall payoffs then it is clear that permanently marking oneself out as exclusive to a particular group is a differentially costly persistent signal: inexpensive for the genuinely committed, dangerous for those likely to leave the community (or fall out with them). But in this case the community need not have a reliable defection-punishment at all, because bad consequences for defection are the result of general outgroup hostility and the signals themselves.

Viewed as temporally extended modelling targets, punitive cultural practices and persistent, group-specific markers like tattoos and ritual scarification can therefore form the moving parts of differentially costly signalling systems. Technically, neither template exactly maps the classic handicap model, as signal cost is vulnerable to community/receiver response (e.g. in ‘punishment’, signalling is only costly for low types who are actually accepted), however in both cases the crucial free-rider game path (being low, signalling high, treated as a high) is made prohibitively costly. More importantly for current purposes, neither interpretation is entirely free-standing in the idealised, maximally parsimonious sense, as they rely on various cultural practices (punishment, parochial altruism) the origins and stability of which call for additional explanation.

6.4.3. Differential benefit and other templates

There is a relatively simple differential cost-benefit template which avoids such complications, and the differential-cost model altogether: differential benefit signalling.

Imagine two individuals considering initiation into a local community, call them Flaky and Staunch. The community is good at productively combining the labours of its members and fairly distributing the benefits (making it attractive to join), but there are significant transaction costs for absorbing new members, and while Staunch is enthusiastically attracted to the group

and its way of doing things, Flaky is far less committed and 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 Flaky has a reasonable risk of messing up at some point, or just wanting out. So, while Staunch can look forward to staying in the community for life (if admitted), Flaky's future in it has a comparatively short half-life. This means that the total future benefits for Staunch will be significantly greater than those for Flaky, but the net benefits for the community would also be different: Staunch would be an asset but investing in Flaky is less likely to pay off (for similar reasoning with regard to the evolution of guilt, see (O'Connor 2016; Rosenstock and O'Connor 2018))

This straight-forwardly fits the sort of strategic situations we are interested in, but there is an obvious way to turn this into a payoff structure that supports a signalling system: set a fixed, up-front cost for entry that satisfies condition [4.3], i.e. one that exceeds the benefit that a fly-by-nighter like Flaky would accrue before moving on. With enough time & experience for strategies to adapt to this costs-benefits regime, paying such an entry fee can plausibly evolve into a reliable signal of commitment.

This template arguably performs quite well with respect to the ideal explanatory virtues of signalling theory. It provides a clear link between commitment level and signalling strategy in equilibrium, since commitment directly determines sender payoffs. Because differential benefits are doing the strategic work, it predicts signals of significant fixed cost but arbitrary form: rituals, demands on time or resources, restrictions/requirements on behaviour, dress, and so-forth that can vary widely between communities. No psychological biases, exogenous constraints, punitive community norms, or other evolutionary pressures are required in principle. A community-erected 'paywall' is also 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, for purposes of clarifying the literature, the general differential-benefit template is also similar to Ianacconne's economic analysis of why strict churches are strong (Iannaccone 1992; 1994), allowing this view to be positioned as a differential cost-benefit fakeable signalling model.

Other temporally extended assignments are possible which mix up the order of signal, cost, and benefit. For a final example consider an 'investment signalling' or 'wise elder' 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 signalling commitment to local groups in a way that is far less expensive (at least in terms of opportunity cost) if you're actually interested in belonging to just a small number – or one – of them. By being sensitive to these pre-paid signals, local communities can again discriminate between the genuinely (parochially) prosocial, and those who will to look for other options when the going gets tough.

6.4.4. Other possibilities

Of course, there even more modelling options available here. One possibility that was foreshadowed in chapter one, and which will be pursued further in chapters seven and eight, is that signalling strategies might have originated in less competitive strategic situations and environments. Under the right conditions, and because of constraints in trait specificity and/or evaluability, costless signalling à la the David Lewis signalling game might evolve for use in common-interest cases, but then 'bleed over' into environments more dominated by the prisoner's dilemma-style conflicting interests. But this requires further discussion that will be left to later.

Vulnerability signalling should likewise be considered as a modelling option, though the operation of the model is different and would require more interpretation. In one sense, we can easily see something like a vulnerability 'signal' in submission rituals. Like a subordinate wolf baring its neck to the alpha male, a ritualistic submission to an authority (e.g. deep bowing or allowing them to place their foot upon one's neck) is only costly if the receiver decides to reject the peace offering. But it is harder to recognise the rest of the vulnerability game payoff structure with respect to the other distributions of costs and benefits here. Also highly speculative is the projection of this game onto the public presentation of a religious authority (e.g. a holy man), whose schtick and promises are rewarded by community acceptance, but could be very costly if they were to turn on him as a fake. The problem with this interpretation is that it is hard to see what it would be for a holy man to *not* be a fake. The actual proportion of 'high types' would be 0, so unless we reconceptualise what is meant by 'high type' (e.g. by building in a subjective perception of benefit for the *receivers*, perhaps based on how good and falsely comforting a fake the holy man is) this will not work either.

This is not to say that vulnerability signalling cannot be linked to the cooperation problem or be placed in a religious context. Indeed, it would be surprising if it could not (as the payoff structure difference between it and intrinsic differential cost signalling are minimal). There may be possibilities for example with regards to temporally extended interpretations, perhaps modifying some of the cases considered above. However the point has largely already been made (that simple models can have many complex interpretations) and pursuing a more complex model in this way will have diminishing marginal returns for current purposes.

6.4.5. Summarising the options

Time to wrap this up. All the previous caveats about the difficulty of fitting models to real life cases should be kept in mind and updated when considering these more fleshed-out interpretations. For example, taking part in an intense ritual might generate commitment via our affective machinery, rather than displaying pre-existing commitment, and so something that seems to fit a Frankian hard-to-fake signal template (and produce broadly similar results) might be due to some other causal mechanism entirely. Likewise, these templates are by no means an exhaustive list, and variations on them (as well as entirely different signalling interpretations) are of course possible. With such caveats in mind, Table 6-1 summarises the seven that I have discussed.

Table 6-1: Seven application templates for models of religious signalling

Template	*Signalling model*	*Context & application*	*Signal cost/honesty mechanism*	*Special requirements*
Cognitive load (Sterelny)	Hard-to-fake/index	One-off/PAYG (dyadic interaction)	Differential efficacy cost (cognitive effort) prohibiting false signal production	• Complex commitment signal conventions • Relevant human cognitive limitations
Emotional display (Frank)	Hard-to-fake/index	One-off/PAYG (ritual setting)	Undifferentiated efficacy cost (pain, time, discomfort) prohibiting convincing false signal	• Affectively/physically demanding ritual conventions • Relevant limitations in sustained affective 'acting ability'
Subjective cost (Sosis)	'Hybrid' fakeable/handicap: subjective intrinsic differential cost	One-off/PAYG (ritual setting)	Undifferentiated signal costs made subjectively differential, directing behaviour as though strategic	• Perception bias correlated with both religiosity & prosociality • Cooperation benefits must overwhelm actual signal cost in evolutionary dynamics
Punishment	Fakeable/handicap: intrinsic differential cost	Extended/pay-later (initiation or maintenance)	Undifferentiated signal with delayed differential strategic cost (punishment on defection)	• Reliable punishment of defectors
Bridge-burning	Fakeable/handicap: intrinsic differential cost	Extended/pay-later (persistent initiation markers)	Persistent signal with delayed differential strategic cost (3rd party rejection on defection)	• Group-specific commitment conventions • Rejection based on rival commitments
Investment	Fakeable/handicap: intrinsic differential cost	Extended/pre-pay (investment for later signalling)	Differential opportunity cost of learning investment (committed types only need invest once).	• Group-specific conventions of learnedness & respect • Significant costs for learnedness investment
Paywall	Fakeable/handicap: intrinsic differential benefit	Extended/earn-later (entry requirements for cooperative community)	Undifferentiated signal/entry cost, but delayed differential accrual of cooperative benefits	• High signal/entry cost

The options here vary along several dimensions: signalling models, cost types and distributions, the temporal ordering of signal, cost, and benefit, and special requirements for a working RST interpretation. Each of these also vary in terms of their explanatory virtues, as discussed in the body text.

The recurring point to bear in mind is that they are *not* mutually exclusive, nor do they exclude other (non-signalling system) causal mechanisms. Many of them might be complimentary. So, if the targets of selection (the various tropes and contents of religious ritual and practice) are well-differentiated enough, then various signalling mechanisms might plausibly feature in explanatory causal complexes at the local community level (that we might even recognise as religions).

More generally, the scientific seriousness of RST will depend on how well its various moving parts are defined and combined: formal models and the mapping of them to real-world features e.g. actual costs and benefits (of the same currency), signal constraints, and relevant update/evolutionary mechanisms. The reader may baulk at the level of detail being recommended, but it is useful at the very least to acknowledge that such levels exist and can (in principle) become scientifically significant.

6.5. Where we are up to

Time then to summarise what has and has not been argued over this middle part of the dissertation.

In the third chapter I argued for a payoff-driven, co-evolutionary reading of RST: religious signals are signals of group commitment which (as long as they are more honest than not) allow cooperative societies to stabilise and grow. The formulation of core RST as outlined is admittedly a narrow one, and possible to disagree with. But I would argue that the formulation here isolates an important explanatory natural kind; offering a more fine-grained comparison of different forms of 'signalling' explanation. This might potentially assist to better carve up 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 fourth chapter I argued for a Lewisian understanding of signalling systems as the basis for honest signalling in RST, and broadly outlined the different abstract ways of ensuring commitment signals are more honest than not. Costs and benefits work differently within different models, and these and other differences are details that can have evolutionary significance. Chapter five continued this with a more formal slant, and justified the focus on making costs and benefits commensurate. I argued for a pluralist, mechanistic approach to religious signalling, and accommodation and complementarity with regards to other explanatory mechanisms, and overall modelling strategies. I also further developed the cost-benefit 'currency' problem for applying signalling models to real-world phenomena, which culminated in the current chapter with discussions of how this is acknowledged in the RST literature, and with respect to evolutionary change. I argued that these distinctions and theoretical requirements are pressing. Nevertheless, the multiple avenues for signalling models make RST seem more plausible when these complex concerns can be addressed (or bracketed of), and I outlined several empirically distinct ways this might be. This concludes a general

progressive arc over these four chapters, from abstract conception of RST to more detailed outlining of its scope for application.

In following chapters I will return to a few specific issues that have arisen along the way, which (I shall argue) are tractable via formal methods.

7. What formal modelling can add

I have argued that it is both possible and desirable to apply signalling theory in a formally detailed way to the evolution of religious signalling and cooperation; to level of detail that signalling models and approximate payoff structures are specified. At this level of detail, we can potentially test the difference between different models of signalling empirically, based on the differing predictions and implications of the models in question. That is not part of the present project. But it is also possible to probe the models more directly via computational methods, and this is the route that this chapter begins to explore.

Computational methods have great utility here because some of a model's implications will be non-obvious. Grasping the formal structure of an evolutionary model is typically insufficient for understanding all the ways in which a system governed by that formal structure might behave. Running computational simulations of these models can therefore uncover regularities or possibilities that are surprising or unanticipated. To this extent then, such simulations can be seen as experiments conducted on the model system, epistemically akin to conducting experiments on a model organism when investigating disease (Parke 2014). If these results are compared to how the putative real-world targets behave, we can use them to test how well the models really fit their intended target. But via simulations we can also probe model behaviours whose analogues in the target system are not readily discoverable otherwise – in effect they elucidate the predictions of any theory which asserts the accuracy of the model. One recent example is the generation of massive libraries of synthesisable chemical compounds (and estimates of their characteristics), by the combinatorial simulation of chemical reactions based on validated reaction models. This allows far more molecules than could ever realistically be synthesised in the lab to be extrapolated out of existing models and triaged (again, in simulation) for investigation against drug targets (Lyu et al. 2019). In other words, simulations can test theories or flesh out their predictions and commitments, depending on how they are used.

Developed in the way it has been over the last few chapters, the RST approach is ripe for such testing. In particular, simulations can help us assess how good an explanation it really is. As far as we know thus far, it is entirely possible that simulations could demonstrate that RST is a complete non-starter. For example, for the sort of applications templates discussed in the previous chapter, it could be the case that none of the signalling models appropriate to them

generate robust basins of attraction for communicative equilibria, within contextually plausible ranges of costs, benefits and other parameters. On the other hand, it might be that specific, contextually plausible deviations away from the simpler, idealised models can sometimes produce a much higher likelihood of signalling, or its co-evolution with cooperation. If that were true, the case for RST might be greatly improved.

7.1. Simulations as evidence and prediction

First though, the dual role of simulations I alluded to deserves some unpacking. Indeed, the literature on models has tended to look for a single, unified role for them in the style of theory testing and prediction, rather than theory exploration.

One recent, compelling treatise is Michael Weisberg's *"Simulation and similarity: using models to understand the world"* (Weisberg 2013). In this book, Weisberg identifies three kinds of scientific models: concrete (where the model structure is an actual structure existing "in the world"), mathematical (the model is a mathematical structure, such as an equation), and computational, where the model structure is composed by sets of procedures, states, and transitions. Each model structure stands in interpretive relations to target systems (real or imagined) and thereby represent them: "interpretations tell us what the model is about and set up the relations of denotation between models and their intended targets" (Weisberg 2013, 15). Each of these models therefore provide the opportunity for 'experimentation' (loosely defined) that via the interpretation can shed light on the target system: running actual physical experiments on concrete models, analysing the mathematical implications of the mathematical structures, and conducting simulation runs of computational models.

How does computational modelling of signalling (ala Skyrms et al) figure into this framework? Recall the two components to the evolutionary models discussed so far. One is the payoff structure that defines the game in question: the series of moves the players can make and their payoffs conditional on those moves. The game structure itself is static though, and the evolutionarily stable strategies derivable from it via analytical methods only show that a corresponding population would be uninvadable, not that it is likely to evolve. As (S. M. Huttegger and Zollman 2013) show, there are many cases in which static evolutionarily stable strategies (ESS) provides few (or even misleading) insights into which of the equilibria are more likely to represent the equilibrium states of an actual evolving population. Populations can be modelled as highly abstract proportions of strategies (a so-called 'continuous'

population), or with agents represented individually and able to interact with one another (‘agent-based’). In conjunction with appropriately defined evolutionary dynamics, it is often possible to give an analytic analysis of how we would expect populations to evolve under the conditions of the game. But this often also justifies computational modelling: simulating various populations of agents ‘playing’ the signalling games, with the replicator dynamics (or some other dynamics sensitive to relative payoffs) governing how strategies within the population evolve over time. In Weisberg’s terms, the simulation involves computationally tractable representations of at least three kinds of mathematical models: the signalling game itself, the dynamics, and the population. The overall computational model results from how these components are combined, and it is constituted by the sets of population states and procedures carried out on them (in accordance to the signalling game structure and update rules).

Of course, simulations of population states and transitions are not representations of actual evolutionary histories, in the sense of mapping to the ways that the target system and its traits of interest might have actually evolved. It would be stretching credulity to go fishing for a combination of a) a simply defined population of signalling strategies such that it approximated some ancestral human state, and b) a payoff structure approximating a signalling game that became available to those ancient humans, such that c) a simulation run using some sort of plausible evolutionary dynamics ends up approximating the actual course of human evolution.

In any case this is not the work that these models are generally put to. The standard strategy instead is *robustness analysis*: the analysis of evolutionarily stable strategies and their basins of attraction across a range of parameter values and (especially) initial population states. The targets of such inquiry are the games and strategies themselves as mathematical objects, ignoring many assumptions about plausible or useful assignments of them (for example, most of the random initial population states of a robustness analysis would never correspond to any real population). In some cases, such analysis is analytically tractable. Computationally though, the method is the repeated application of the game and evolutionary dynamics to randomised populations and counting the proportion of populations which end up in the equilibrium states of interest. Less commonly, some modelers do indeed take further interest in whether a computational model demonstrates the stabilisation of a behaviour of interest in a population, or its emergence from a relevant ‘ground state’ where the behaviour is absent, perhaps representing an ancestral state. With respect to signalling or the cooperation problem, a

‘ground-up’ simulation would start with no (or low) levels of signalling or cooperation, and test for ways in which they might be bootstrapped into existence. Broadly speaking though, robustness analyses and ground-up approaches look at the in-principle evolutionary pressures that certain strategic situations and mechanisms might account for. Particular outcomes might demonstrate *proof of concept* for their potential evolutionary significance for stabilisation (in the case of robustness testing) or inception (for ground up simulations). But in any case, the proper investigative targets are the general shape and relevance of potential evolutionary pressures that different forms of signalling can generate; rather than an impossibly precise, general, and realistic evolutionary narrative. So, while modelling work can produce subtle and surprising results, we should not over-interpret simulation results. This is especially so in the context of the assumption of causal complexes of which signalling mechanisms would only reasonably figure in as components (as discussed in previous chapters).

It is in this sense that we can see this simulation work as being more about providing evidence about the theory than evidence about the world (especially if the theory is otherwise under-specified). Theories such as RST are hypotheses about the causal role of signalling mechanisms in generating the phenomena we observe. The formal investigation of signalling models gives us information about what sort of phenomena are likely to be generated, and under what conditions. It increases the epistemic specificity of these theories, allowing them to be more precisely refined, calibrated, and tested against empirical investigations into the real-world phenomena.

But engagement with observational or experimental science on the target system is a further step, which implies a further ‘strategic’ decision-point regarding what modelling work can be used for, that depends on the epistemic status of all the various moving parts. For some of the models that Weisberg discusses, for example a large ‘concrete’ model of the hydrology of the San Francisco bay area (coincidentally, mostly made of actual concrete), the work being done is very much from model to world. That is to say, the observable behaviour of the model is combined with a certain confidence in the relevant similarity between the model and the target system to generate predictions about how the target system will behave. That is because the behaviour of the world is the least certain, and the target of the enquiry.

Not so in our context. The goal here is to find a good explanation (or component of an explanation) for the evolution of human religions and cooperation, not to predict the behaviour

of the target system. In the RST context, we know a fair bit about how the target system ended up, but not how it got that way. The uncertainties are instead about mapping model to target, which we can mitigate to the extent we can find (especially novel) model-target similarities. If the model behaves in a way such that it looks like a real-world causal analogue would help produce what we see in the target system, then that is (perhaps) some degree of confirming evidence that this might be an appropriate model. If it cannot be made to produce such results under plausible parameterisation, then that is a presumptive falsification of that particular configuration of the theory (i.e. with the signalling model that was used, applied in the way it was).

There is some dispute about how exactly to characterise the 'evidence' that formal methods provide, though for our purposes only the language really matters. Weisberg describes robustness analysis as providing 'weak confirmation' (Weisberg 2006), whereas (Forber 2010) describes it instead as 'constraining biological possibility'. This latter description is probably the more apt, though I would again place a heavier emphasis on the model/theory side of biological possibility. What simulations of reasonable signalling models can do is map out the limits of explanatory possibility for religious signalling as an evolutionary explanation. The modest characterisation then is that model-based RST hypotheses are 'how possibly' explanations: specifying adequate *candidate* mechanisms. The stronger evidential import of modelling is that we can rank and contextually assess them in comparison to each other and non-signalling explanations. A rigorous approach to RST should posit putative real-world mechanisms which have tight formal analogues (in the sense of being assignments of signalling models). Probing the behaviour of the formal analogues can thereby narrow down the space of the possible, by better fleshing out what causal work the putative mechanisms are ideally capable of doing.

7.2. Under what conditions will signalling and cooperation co-evolve?

In the rest of this chapter I will first review some of the prior relevant literature, and then lay out the motivations for some simple simulation 'experiments' whose results to be described in the following chapter. These will be simulations of signalling evolution where the signalling models and parameters are modified away from their commonly explored, idealised forms, in ways motivated by the previous chapters' discussions of RST.

Recall from the last chapter that there were several open questions for RST, some of which would be tractable by formal methods. The key one being the co-evolutionary question: under what conditions will signalling and cooperation co-evolve? The whole rationale of RST (at least the one under consideration here) is to explain the evolution of religion and cooperation as a package deal. It should be acknowledged up front that this is not generally the concern of the formal modelling literature. Most of the formal literature discussed so far shows how signalling (costly or otherwise) can be both stable and likely in various evolutionary signalling games. However, this is generally with idealised, stipulated payoffs (i.e. specifics of payoff generation ignored or held fixed) and in the RST context payoffs are generated by the benefits of cooperative behaviour. In other words, drawing lessons about religious signalling from most of the general signalling literature is to treat cooperation as an exogenous variable: we can make inferences about how likely signalling is to evolve given certain background levels of cooperation. As we shall see, even when the cooperation literature includes informational features, the ways those features are modelled tend to be narrow and idealised in a way that does not correspond to the detailed signalling forms used in the mainstream signalling literature. But much literature does exist, and it paints a cautious picture for would-be RST boosters.

7.2.1. Tag, green beard, and ethnic marker models

One classic and widely-referenced example here is (Riolo, Cohen, and Axelrod 2001), which demonstrates the evolution of cooperation in an agent-based model via recognition of arbitrary ‘tags’. The agents are randomly paired but decide to cooperate (in the form of making a donation) or not with a probability depending on how similar their tags are – there is no record of prior cooperation available to the agents, and therefore no scope for evolution of cooperation via reciprocity. There are a number of interactions before a payoff-sensitive update rule is applied, with the more successful agents having more offspring who inherit (though not perfectly, because of a ‘mutation’ process) their parent’s tags and tolerances (i.e. their similarity thresholds for donation). In the terminology introduced in the first chapter then, cooperation is stabilised via a formal analogue of kin recognition and a kin selection-like mechanism, with the tags in the model playing the role of a green beard. The cooperation that stabilises is based on the coalescence of ‘dominant clusters’ of cooperating, similarly-tagged agents, which (in broad terms) illustrate an analogue of parochial altruism.

In the present context, this study does not directly address or precisely map the sort of signalling-cooperation co-evolution that RST proposes, and it has other interpretive limitations as well. In the language of signalling theory introduced in chapter four, the tags of Riolo et al begin as loosely associated cues – insofar as weak, accidental correlations between them and are laid down by the initial random distribution (due to small numbers). Via selection, these become fakeable signals correlated with conditional cooperation. Importantly though, the initial donation rates and tolerances of agents start off uniformly distributed among the population (with average tolerance being quite broad) rather than at a low or ancestral state, meaning that the proof of concept it supplies is more of stabilisation than inception. It is also a structural feature of this model that there is no "always defect" strategy made available to the agents: even an agent with a tolerance of zero will still donate to tag-identical agents. And this non-zero 'zero' tolerance is also a 'hard floor', meaning that mutations will tend to perturb it upwards. Overall, this arguably gives the evolution of cooperation an unrealistic advantage in this model (there are no 'anti-social', 'iconoclastic', or 'seek out difference' strategies available either). As demonstrated by (Roberts and Sherratt 2002), simply allowing tolerances in the Riolo et al model to venture into negative territory means that dominant clusters of cooperation are far more susceptible to invasion (by tag-similar free-riders). How significant a difference this makes is a matter of dispute (Riolo, Cohen, and Axelrod 2002). In any case, as commonly found in green beard-like mechanisms (see chapter 1) the tag-based 'dominant clusters' of this model do not last long, with establishment and then invasion by less tolerant/donation-prone free-riders. Sustained levels of cooperation require that agents are able to cycle between different tag values (i.e. available beard 'colours'). This 'chromodynamics' has been highlighted by subsequent formal research which shows the need for tag mutation rates to be many times higher than those for strategy mutation rates (which is not at all realistic) in order to sustain cooperation at a majority level (Hales 2005), and/or the requirement for inheritance of behaviour & tag to be only loosely coupled (Jansen and van Baalen 2006).

A model which uses a similar tag/tolerance mechanism but does not rely on Riolo et al.'s modelling affordances or strongly asymmetric mutation rates is given in (J. P. Bruner 2015b). In this model cooperation is a far more likely result and can be shown to increase from a very parochial state (i.e. a homogeneous population cooperating with only very narrowly similar partners) to one that is both heterogenous and broadly tolerant. The key to this result though is that the payoff structure of the game must be a stag hunt, rather than a prisoner's dilemma. This

coheres with other work on how cheap signals can greatly improve the likelihood of converging on the more valuable of the two stag hunt equilibria (Skyrms 2002), and the dynamic role of 'ethnic markers' in demarcating and coordinating behavioural/normative compatibility (McElreath, Boyd, and Richerson 2003). On the basis of this later simulation work, the authors argue that unconstrained, un-incentivised arbitrary markers do not function to direct altruism between similar agents (because of vulnerability to invasion), but instead function to advertise behaviour types relevant to *coordination* benefits (because there is no incentive to deceive or manipulate). The empirical evidence is also seems to agree with this conclusion, with a large Danish empirical study of the role of accent similarity on behaviour in public goods games showed that similarity of regional accents predicted better performance in coordination games, but not improved cooperation in competitive conditions (N. H. Jensen et al. 2015).

So, while tag-based models might loosely fit the definition of fakeable signalling outlined in chapters 3 and 4, they appear to fit less well with the target systems and hypotheses of RST. Tag-models support cooperation in coordination-like games, and in more competitive strategic situations up to and including stag-hunt payoff structures; this fits with what we should expect from costless signals. But tag models struggle in prisoner's dilemma-style strategic situations, where 'low types' seek to invade and must be excluded for cooperation to be worth-while for a community of receivers. This may be mitigated by high mutation-rates for tag/signal content, but successful bootstrapping of cooperation via this method requires that this content be inherently unstable and subject to a rapid turnover-replacement cycle relative to the rate of strategy change. And of course, this doesn't fit the received picture of religious content very well at all: one of the outstanding questions about religion is why religious content (within a tradition) is relatively highly conserved over hundreds of years, compared to other cultural features of the relevant society.

7.2.2. Public goods as costly signals

This leaves the door open for incentivised or constrained signalling (i.e. differential cost-benefit or index signals) to play a role. One classic paper in this area is (Gintis, Smith, and Bowles 2001), which uses a model more closely resembling the signalling/cooperation co-evolutionary picture that we are after. Like Riolo et al., Gintis et al. use an agent-based model, where no reciprocity is possible, and where players are given the opportunity to make a costly donation. In this case though the donation: i) is a public good, in that the benefit is shared among the population as a whole (but at a net cost to the donor), ii) serves as a costly signal

which the rest of the population may observe (at a small cost), and iii) is followed by a partner-choosing step (for each agent) that results in further costs/benefits being distributed. Determining donation cost (but invisible to the other agents) is an agent's underlying type, 'high' or 'low', and this also determines the benefit of choosing to partner with that agent. High types pay less to make the signal donation[68], but are worth more if chosen as a partner. In their sender role, the agents have the 4 familiar pure strategies of: always donate, never donate, donate if high, or donate if low. In their partner/receiver role, they select potential partnering allies (to choose randomly between) based on any observed donation (or not), or at random. Every time an agent is selected as a partner (more than just once), they receive a fixed benefit, and overall payoffs (via the replicator dynamics) determine how both the type traits and the sender-receiver strategies evolve.

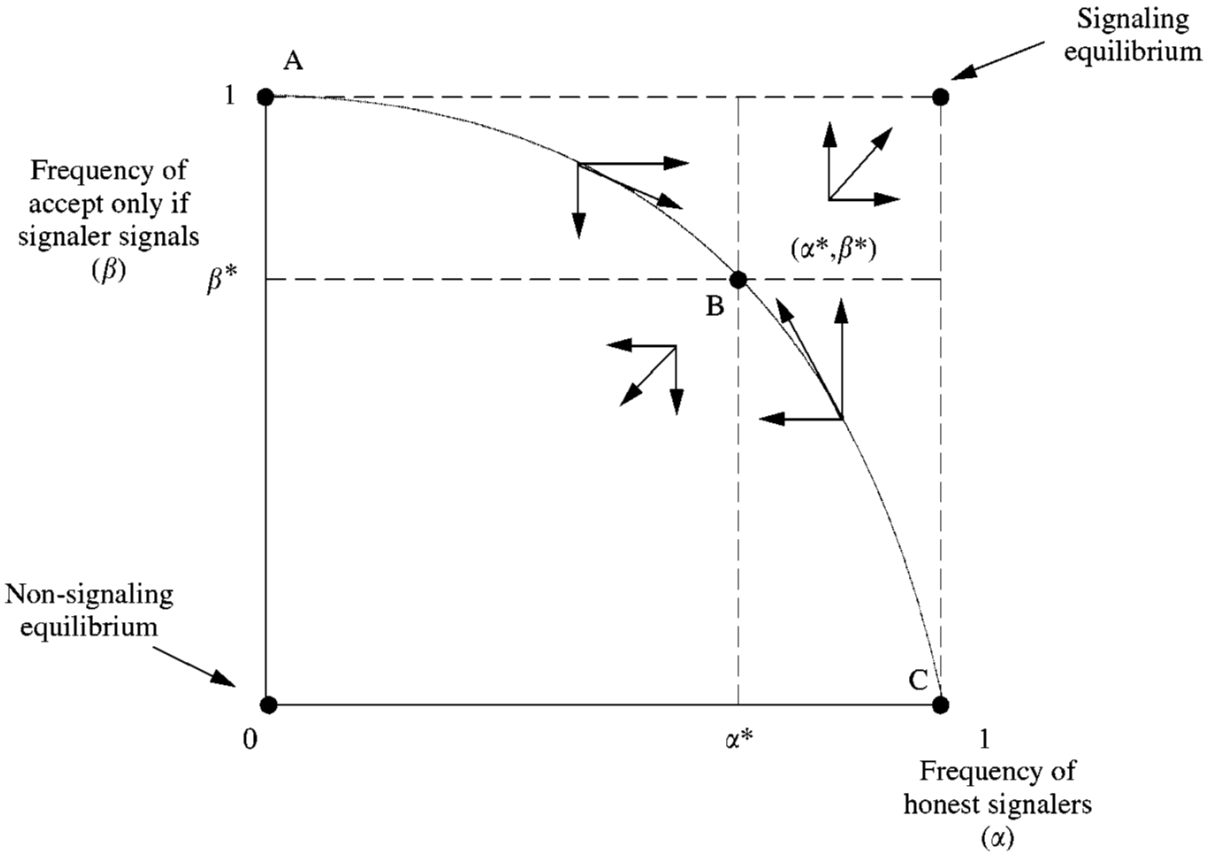
Using detailed analytical methods (i.e. without simulations), Gintis et al. show that 1) there exists a signalling equilibrium which corresponds to the standard costly signalling separating equilibrium (as long as costs and benefits are appropriately aligned), and 2) that under several further conditions (including the proportion of high types being not too small, and the population being sufficiently large), there is a significant basin of attraction for this equilibrium. Honest signalling is fitness-enhancing for the group as a whole, however this doesn't even necessitate the value of the public good 'donation' being net positive, as long as many alliances with high types occur.

There is a lot to unpack here. The result appears to fit the RST narrative: differentially costly public displays support the evolution of honest signalling. The authors also see their model as a general framework for modelling a variety of social phenomena, including sharing of resources, defending the group from attack, or costly prosocial punishment of defectors. Without getting into too much detail though, the results here also speak more to stabilisation than inception. This is because there is no evolutionary pathway from low rates of signalling to a signalling equilibrium. Figure 7-1 shows a phase diagram reproduced from (Gintis, Smith, and Bowles 2001), which illustrates the evolutionary dynamics of the sender and receiver strategies for the signalling equilibrium (again, given conditions specified in the paper). The x-y parameters here are the prevalence of the conditional sender strategy, and the prevalence of

[68] It is worth noting that the authors see their model as generalising to what I have called differential cost-benefit signalling, in that differential benefit instead of differential cost can be expected to demonstrate the same results.

the matching conditional response. Arrows from points on the diagram show evolutionary trajectories from that point in parameter space. The diagram therefore shows that the signalling equilibrium will only be reached from regions above the A-B-C frontier curve, i.e. where one or both parameters are sufficiently high (this is its basin of attraction). For a point below the frontier, the non-signalling equilibrium is inevitable. The exact shape of the frontier between the two basins of attraction (and the values of α* and β* where its saddle-point is located) will depend on the various cost and benefit parameters, and the prevalence of cooperative types in the population.

Figure 7-1: Phase diagram from (Gintis et al. 2001). Points below the ridge line ABC are in the basin of attraction of the non-signalling equilibrium.



The evolutionary limitations this result implies are acknowledged by the authors, who discuss a more-or-less standard set of further reasons why prosocial signals might be favoured (many of which have already been touched on here). In the RST context, this is in effect an acknowledgement that this form of signalling model is best seen as a way of stabilising and driving to fixation signalling-like behaviours which must originate by other means.

And even this main result does not consider co-evolution of high types with the signal for them. A necessary condition for the stability of the signalling equilibrium here not just that the proportion of high types is not too low, but also that it is not too high. This should not be surprising: signals (especially costly ones) are only valuable if they convey valuable information, and they have no such value in a society where *everyone* is a high type. Signalling "I am a high type" would be wasteful. Waiting for such a signal before acting (rather than just picking a partner at random) would only disadvantage the receiver. Indeed, one of the lessons drawn in (Skyrms 2002) is that despite facilitating convergence on the optimal stag hunt equilibrium, the informational utility of signalling is strictly transitory (i.e. during the process of convergence only), and it becomes a useless vestige once convergence is complete. A signal that proposition p is true has no significant evolutionary advantage in environments where p is always true. In the case of the (Gintis, Smith, and Bowles 2001) model, the lesson is that for any configuration of this model, there will be a low-to-high range for the proportion *p* of cooperative high social types among the population, within which the signalling equilibrium is stable. The lower bound of this band of stability is not 0, and the upper bound is not 1. The good news that Gintis et al. demonstrate analytically is that the evolutionary dynamics for *p* are such that it will have an equilibrium point within that band, meaning that the proportion of high types does not rise high enough to destabilise signalling. But this is of course achieved by via cooperation being held back from fixation.

This should temper expectations for a 'runaway process' narrative, whereby the co-evolutionary interaction of signalling and cooperation drive each other from next-to-nothing all the way to fixation. In terms of scientific promise, whether stopping short of cooperative fixation is seen as a bug or a feature will in part depend on assessments of the distribution of this social quality within actual religious populations. But in any case, we can't expect this kind of signalling (in the form this model represents) to provide much help with the inception of signalling and cooperation. Both need to be present to some significant degree for the co-evolutionary dynamics and balance of payoffs to gain positive traction, and indeed below certain levels it is predicted that they would supress one another. Even under ideal conditions, something else must supply enough variation and prevalence of signalling-like behaviours before evolutionary signalling proper can drive them to fixation, and cooperation to its optimum.

7.2.3. Possibilities and speculations

The formal results here are patchy but instructive[69]. As expected from the general discussion of signalling theory, the work on tag-based signals demonstrated that cheap, fakeable signals are not able to stabilize cooperation without either deviating away from the general criteria of the cooperation problem, or from a useful application to human religion and ritual. They *can* play a role in keeping cooperation levels elevated if signals rapidly shift or mutate; and can drive up cooperation from low levels in coordination or stag hunt-style strategic situations. But neither of these fallbacks seem to fit RST interpretive paradigm. The work on costly fakeable signals on the other hand shows that signalling can be stabilised, perhaps along with a high (though not universal) level of social cooperation, but it needs both signalling strategies and cooperation to be already present in the population, at levels that cast doubt on it as an evolutionary explanation for their inception. This might be enough though, if we think of these forms of signalling as coming online to maintain cooperation in social environments as they (for example) become larger and/or less transparent, or as strategic environments change, signal forms become available, and so-forth. Such a 'meandering route' into a robust signalling basin of attraction is not an entirely ad hoc suggestion and arguably quite in keeping with narrative suggestions in the literature.

Of course, the degrees of freedom for modelling decisions are such that these doubts are not comprehensive or crushing. For every specific negative result, there will be many 'but what if…' responses which point to alternative lines of inquiry, in both modelling and interpretation. For example, Gintis et al. argue that their results should generalise across differential cost and differential benefit models. But the costs/benefits they describe are intrinsic costs, and they look to pure strategies generating separating equilibria. Would a vulnerability cost model behave differently? Would the picture be different if hybrid equilibria were considered? Would

[69] Other, less well-examined modelling results are also potentially relevant, for example (Müller and von Wangenheim 2016) claim to have found chaotically stable cycles of cooperation and signalling, via what they call a *transitional* equilibrium, which is "based on the dynamic interplay of separating, semi-pooling, and pooling equilibria [and] can stabilize a population state characterized by heterogeneity with respect to all three dimensions: preferences, behavior, and signaling" (Müller and von Wangenheim 2016, 20). More recent signal-cooperation co-evolutionary work builds in further factors such as group selection e.g. (Scheuring 2010). For our purposes though there are diminishing marginal returns for surveying such work, and general lessons can still be drawn in the absence of doing so.

simulations based on this model add nuance to the analytical results? There is scope here for formal work to address these questions.

But formal analysis can suggest explanatory possibilities as well as critically examine them. A formally aware ‘what if’ narrative might go as follows. What if there was a pre-religious ‘ancestral’ cultural state with a chaotic roil of spiritual tropes, stories, and ritualistic behaviours coming and going (and with their own origin story), but some of these were upregulated (exapted) along with cooperation by (contingently) coming to serve the function of tags or ethnic markers to facilitate a level of mutually beneficial social coordination (perhaps leveraging pre-existing kin-recognition instincts). This tended to happen when times were good and when defection was more of an annoyance than a life-threatening betrayal or assault – the economic and strategic context of the population meant that more of interactions were in coordination or stag-hunt-style strategic situations than prisoner’s dilemmas. When conditions get worse, the buzz of proto-religious complexity and woolly commonality tended to die down again (sources of variability would depend on evolutionary time frame, but might be as short term as seasonal changes in conditions and economic/resource base, periodic switching between migratory modes of living, or stochastic ‘bad years’). However, sometimes the changes in circumstance would also alter the economic context within this pre-cooked stew of signalling tropes and/or cooperative behaviour – they became differentially costly or beneficial in such a way that a viable fakeable signalling mechanism was constituted (e.g. loosing access to a ready supply of ochre or some other symbolic resource, requiring the calling-in of favours). If other ingredients and quantities within the stew were favourable (i.e. sufficient levels of conditional strategies and high-quality social types) then a more robust signalling mechanism might take root and stabilise them. Such a scenario does not seem to be ruled out by the results here.

Obviously, much more work would have to go in to turning ‘what if’ speculations into ‘how possibly’ explanations, or more grounded lineage explanations. And the context of inquiry here is artificially narrow: it is implausible to consider these two mechanisms working together in isolation from any of a wide variety of potentially explanatory non-signalling mechanisms and causal influences.

Once we attempt to serious link simulations to their target, another set of moving parts and associated layers of detail and complexity for fleshing out RST comes into view. We should

have plausible evolutionary processes at work, on signalling games that are a plausible fit for the social-religious contexts in question, with plausible interpretations of cost and benefit, and in the right order or sequence so that we remain within parameter ranges where the dynamics of those models push behaviour in the right direction. Though the basic idea of religious signalling theory is relatively straight-forward, fully specifying it is anything but. Again though, the required level of complexity here depends on the level of detail or generality that is being sought after. And there is at least potential for being able to discover general trends about what sort of factors might promote or supress the evolution of prosocial signalling.

7.3. Asymmetry and deviation from ideal signal form games

In the previous section we saw a way of deviating away from idealised, symmetric signalling models which allowed an improvement in prosocial signalling: increased mutation rates for signal content in tag-based signals. The applicability of tag-based models to RST is doubtful. But this sets up two further questions to consider in this chapter:

1) Are there any deviations away from ideal, 'symmetric' signalling models which *are* a good fit for RST and would also act to upregulate the likelihood of signalling and/or cooperation?
2) Can formal modelling say anything about which of the RST-appropriate signal forms (fakeable or unfakeable) are more evolutionarily significant?

The discussion of these questions will set up the original modelling work in the final chapter.

7.3.1. De-idealisation and symmetry breaking in the David Lewis signalling game

Recent work on Lewis signalling games (see 5.1.1) provides a template to survey the many ways in which such de-idealizations could occur, and the robustness of signalling in the face of them. Some deviations from the standard Lewis signalling game include i) more and varied states of the world, ii) the possibility of observational error or signal error, iii) noisy signals, iv) the introduction of a partial conflict of interest between senders and receivers, v) the reception of more than one signal, and so on. Many such concerns are dealt with favourably in (Skyrms 2010), and in work by others. For example, (J. Bruner et al. 2014) generalize beyond the 2x2x2 signalling game idealisation (illustrating partial-pooling equilibria with 3-signal games) and also provide experimental confirmation of the expected convergence toward coordination behaviour in human subjects, using the methodology of experimental economics. In (Godfrey-Smith and Martínez 2013) and (Martínez and Godfrey-Smith 2016) signalling

games of common interest and conflicting interest are mixed, quantified and compared, demonstrating an increasing monotonic relationship between degree of common interest and likelihood of populations reaching signalling equilibria.

Another particularly important set of studies (for our purposes) show that varying the world-state probability in simple Lewis signalling games can reduce their basin of attraction as 'nature' exhibits a bias. In the standard 2x2x2 David Lewis signalling game we assume that world states w_1 and w_2 are equally likely, i.e. that Nature, treated as a player in the game, makes the first move to decide between the two world states but does so with even, 50-50 probability. Under this assumption, the basin of attraction for separating equilibria is 100%: every initial mix of strategies will converge to a separating equilibrium given positively payoff sensitive evolutionary dynamics. However, if nature exhibits a bias and the probabilities of the world being in one state or the other are not equal, then there exists an evolutionarily significant pooling equilibrium in which no communication occurs between sender and receiver (Huttegger 2007; Pawlowitsch 2008). This happens for similar reasons as observed previously: the closer the probability of w_1 ($P|w_1|$) is to 1, the less value there is in investing in a signalling system for it (i.e. for the Gintis et al. model, $P|w_1|$ equates to p). For some initial distributions of strategies, and where $P|w_1|$ is high, an unconditional, "always act as though w_1 is true" response will have a significant initial fitness advantage and may (depending on the evolutionary dynamics) be driven to fixation before an overall conditional sender strategy emerges to incentivise a conditional response. In other words, in some conditions it is easier and more profitable just to guess, and this results in a pooling equilibrium.

This shows that the results of idealised, symmetrical signalling models can be deceptively 'clean'. For example, in imagining how a signalling system might evolve, there can be little empirical rationale for presupposing an approximately even chance of the trait/state of interest to be signalled for. If a Lewis-style signalling system *does* evolve, then perhaps that is a reasonable inference (because otherwise the it would have been less likely to evolve). For example, just as a Gintis et al.-style model requires a significant initial level of prosociality to serve as subject matter for signalling, a Lewis/Skyrms-style explanation of the evolution of language (for example) would predict that the first communicative utterances would have been about states of the world which were a) of practical importance to both sender and receiver and b) neither rare nor universal. But in the context of looking to signalling as a putative explanatory mechanism, independent estimates of the ancestral state of $P|w_1|$ should be allowed

to speak for or against that explanation – model-informed ideal conditions should not be assumed.

7.3.2. Sender-receiver asymmetry[70]

Another symmetry that is commonly 'baked in' to evolutionary signalling models is that between senders and receivers, in terms of their available options and evolutionary responsiveness. Senders and receivers (in the evolutionary treatment of such games) are two populations of highly abstract and constrained agency roles: all that senders do upon observing the state of the world is send a signal, and the receivers chose to act as though the world is in one of the two possible sender-observable states.

However, in many contexts this can be seen to be an unrealistic idealisation. And in general, of those two roles, it is the restriction on receivers which appears to be particularly unrealistic. By design, the 2x2x2 Lewis signalling model and its derivatives are constrained regarding the receiver's actions: they are limited to just those acts relevant with the sender's observed world-states. And this of course makes perfect sense as a simplifying idealisation. But I would argue that the degree of idealisation here – the distance between model and target system – is greater for receivers than it is for senders.

Consider the following example, involving two foragers suddenly presented with the opportunity to coordinate. One observes a prey animal at a distance, but in such a location or context that it is inaccessible to her. The second forager is prevented from observing the prey by an intervening obstruction but would be able to acquire it by approaching round from a specific direction. Both would benefit by being able to direct the second forager's actions to that task, but a workable signalling system is called for to achieve that. Given enough experience of such situations together, we can imagine a loose collection of conventions emerging between these two which (through trial and error) come to constitute such a system. But being a sender and a receiver in these contexts is quite different. In this case, it is easy for the first forager to slip into the signalling role and execute it; calling out, whistling, or gesturing to her counterpart. But to play the receiver role the second forager has to actually re-orient her attention (to some degree), rapidly draw some very specific interpretive conclusions about what

[70] Parts of this and the following sub-section reproduce/rework material that was previously published (in C. Brusse and Bruner 2017).

the signal means in this context (e.g. prey is to the east or to the west, rather than a predator etc.), and attempt to engage in appropriate behaviour that will be more involved and expensive relative to the signal.

This illustrates three broad, interrelated points of dissimilarity that can occur between senders and receivers in interpretations like this. One is a broadly *economic* dissimilarity: the signals here are cheap and easy to send, yet the actions available to the receiver are less plausibly interpreted as intrinsically cheap and free of opportunity cost. Securing the animal will take more than whistling to it, and in doing so the receiver might be forgoing other possible actions (in a way that the sender does not). Second is an *epistemic* dissimilarity; the informational states drawn on by sender and receiver are also very different. Any real-life sender's observation of a world state will likely inform their motivations ('we should catch that animal') and dictate a fairly clear course of action (try to direct the other agent's behaviour). But all the receiver gets is a whistle, gesture or other signal which (by stipulation) has no pre-established meaning. The experience of observing a strategically relevant state of the world will generally be richer and more detailed than that of observing a strategically relevant artificial signal. Following on from these, it is also reasonable to consider a *cognitive realisation* dissimilarity. Given the likely differences in informational states, goal-directness, workload and opportunity cost between sender and receiver roles, they are likely to require the use of quite different subsets of the agent's available cognitive resources. In other words, we can expect the mechanisms (cognitive and otherwise) which realise those roles to differ as well, quantitatively and qualitatively. A signal can become a learned, reflexive action, based on a limited repertoire, whereas a response (at least in this example) will involve a far more cognitively involved series of decision points and investments, with a far more open-ended tree of potential outcomes to be narrowed down.

I argue that such dissimilarities imply issues of 'fit' for the standard idealisations of signalling models based on evolutionary game theory. First, there is a *structural symmetry* issue, relating to the payoff structure which defines the signalling game. Strategic asymmetries are likely to exist between senders and receivers, not just in terms of payoff, but also plausible and payoff-relevant actions and strategies. Receivers are likely to have locally reasonable options option to them, other than those relevant to sender-observed states of the world, and their responsiveness to the strategic situation is therefore less satisfactorily modelled by the strictly symmetric payoff structures of standard signalling games.

Second, there is an *evolutionary symmetry* issue, relating to the way that evolutionary dynamics are to be applied to each agent/role and their realiser mechanisms. The usual assumption is that senders and receivers update their strategies in an identical manner, modelled using the same dynamics, operating at the same rate. But the epistemic and (especially) cognitive dissimilarities imply that we should not expect the update-responsiveness between sender and receiver to be equal either. This might be especially true if we are thinking of update by learning in contrast to biological evolution, and with cognitive and epistemic asymmetries in mind. When considering signalling strategies, what we are assuming is that there are realiser mechanisms for those strategies, and these will likely be complexes of i) deliberative actions based in general cognitive capacity, ii) enculturated learned behaviours, and iii) developmentally entrenched capacities and other traits, mediated by inheritance mechanisms ranging from the genetic to the purely cultural. It is therefore possible for the evolutionary responsiveness of such strategies to vary dramatically, depending on the exact blend of cognitive mechanisms that they call upon.

In part, this second symmetry issue is motivated by the earlier observation that there are several different ways in which the evolutionary dynamics for sender and receiver strategies might be reified: e.g. biological, cultural, or economic. That is to say, by natural selection of biologically encoded strategies, cultural/social transmission of learned strategies (with selective modification by social learning), or rational update of volitionally calculated strategies (based on cost-benefit assessment). Even assuming payoff-sensitivity (as opposed to purely imitative or other mechanisms), these are often modelled via different evolutionary dynamics algorithms to better approximate the assumed update mechanism.

Strictly speaking, this should not make much difference for straight-forward modelling purposes (again, providing that payoff-sensitivity is the primary driver of the updating, as we are assuming in the context of RST). Evolutionary models are just models of change, and there is considerable freedom in the assignment of model behaviour to target processes. The iterations of a computational model could be assigned to the target system so as to approximate either the reproductive generations of the real population, or an arbitrary time-step after which a matching proportion of individuals in a population are expected to have rationally updated their strategies. The same evolutionary model could adequately model either process.

But what is arguably more significant about the processes being modelled is the difference in timescales that they operate over, i.e. their time-responsiveness to evolutionary pressures. For example, imagine there is a sender strategy with confers benefits on senders that correspond to a 10% improvement on fitness compared to other strategies. After one generation, natural selection should have improved the populational share of that strategy a modest amount. But social learning might be far more sensitive than this: if the individuals deploying this strategy become seen as models for emulation by learners, the strategy might be converged on in only a few generations. And if the adoption of strategies is conscious or volitional, and optimal strategies are actively searched out and switched between multiple times within a single lifetime, the convergence could be even more rapid. So, differences in kind with respect to update mechanisms for senders and receivers (as seems likely) could imply very different evolutionary timescales.

7.3.3. Asymmetric signalling games in the literature and for RST

The structural symmetry issue (as I have argued for it) parallels one of the worries that (Sterelny 2012b) articulates about the general methodology of Skyrms regarding the David Lewis-style signalling models, in (Skyrms 2010). Sterelny asks whether the availability of 'third options' on the part of the receiver might undermine the evolution of signalling even when these third options are less valuable than the payoff for successful coordination. As part of a discussion of animal threat responses, he labels this a 'hedgehog' strategy – because it provides the agent with an action that pays off modestly regardless of the state of the world. To use his example, hedgehogs often roll into a ball in response to predators. This is a stark contrast to the more sophisticated behaviour of vervets, who have specific responses to specific threats. Yet the optimal response a vervet takes to one threat – climb a tree when confronted by a leopard – may lead to total disaster when used in response to another threat, such as an eagle. The hedgehog avoids such risks by 'hedging'; using an unconditional strategy that does not provide an optimal response to specific situations, but an adequate one to most, avoiding the risk of complete disaster. Translated into signalling games and our forager example, the receiver would be deploying a hedgehog strategy if she ignores the signal and the prey animal which it represents, and instead focused on securing a lower-value resource that she would otherwise forgo by attempting to respond. And of course, this is more general than Lewis signalling games. For example, the 'hedgehog' strategy here is in many ways analogous to the risk dominant 'hare' response in stag hunt games. Playing hare instead of stag allows the agent to

avoid exploitation and disaster, but only guarantees the individual a mediocre payoff. Something like these hedgehog strategies is a plausible departure from the idealization of the baseline Lewis signalling game and better captures the demandingness of the receiver role. The question is whether (as Sterelny suspects) including hedgehog strategies might undermine the evolutionary robustness of signalling systems.

In the religious signalling context, the hedgehog strategy is open to several interpretations. The receiver in RST is an individual or community, who has the choice between treating the sender as a trustworthy, socially valuable high type, or to exclude them. While these look like the only choices of action, this is not necessarily true. For example, it might be that some communities opt to ignore signals and protestations of commitment in some cases, and accept applicants into lesser, provisional roles where they are not trusted with the full responsibilities to either generate communal benefit (or extract from it), but instead are put to work in a way that generates the same level of benefit regardless of the candidate's underlying quality. Think of this as akin to getting new recruits to peel potatoes instead of manning the walls or guarding the children: it is a waste of human resources should they be as brave and upstanding as professed, but reasonable insurance in case they are not. Another concession we should make to reality is that signals are *not* the only mechanisms for assessing the quality of potential interaction partners, and instead relying on 3rd party information would effectively be a hedgehog strategy. As an example to demonstrate this, assume that limited information is available about all new potential interaction partners in a simple 2x2x2 signalling game, which would allow receivers to make an accurate assessment of sender quality (i.e. underlying world-state, w_1 or w_2) 80% of the time. In effect, relying on this information instead of any signals received is an additional strategy with a uniform expected payoff of 0.8 (assuming a normalised 'success' payoff of 1 for choosing correctly). Importantly, it is uniform in the sense that the payoff is *not* sensitive to the probability of the world-state, unlike other unconditional strategies (i.e. always a_1, the best action for w_1, or always a_2, the best for w_2). This means that hedgehog strategies are distinct and potentially interesting both formally and in application to RST[71].

[71] Additionally, hedgehog-strategic information sources might include gossip, which suggests a potential line of inquiry with respect to indirect reciprocal altruism. Given the rationale for RST (from chapter 3) that it improves on indirect reciprocal altruism in larger populations, understanding the robustness of signalling in the face of hedgehog strategies might also help us better understand the expected thresholds for this. E.g. at what point does

The evolutionary symmetry issue also parallels a well-known evolutionary hypothesis: the so-called Red Queen effect. In competitive relationships such as predator-prey or parasite-host, the Red Queen hypothesis states that species will be constantly adapting and evolving in response to one another just to "stay in the same place" (van Valen 1973). Assuming similar evolutionary pressures and responsiveness, there is much change to be predicted but no advantage, as an adaptation on one side will be met by a response from the other. In (R. Dawkins and Krebs 1979) however we see an example of two groups adapting and evolving at different rates in the famous "Life-Dinner" principle. While we expect both predator and prey to adapt to each other, we might expect the prey species to evolve at a faster rate than the predator species due to the different selection pressures exerted on both species: failing to adapt quickly enough for the predator means going hungry for an extra day, while for the prey, failing to adapt means death.

It should be clear though that differences in evolutionary responsiveness are not just a matter of asymmetric payoffs or other unequal selection pressures, as various factors contribute to the capacity to respond to evolutionary pressure. Most trivially, microbial pathogens of humans can evolve many times faster than their hosts simply due to a massive disparity in reproductive timeframes. As a result, their most important evolutionary arms races are largely not against humans, but rather against each other. With competitors on a more equivalent time-scale, evolvability and robustness too can play a role, with respect to the possibilities and pathways for adaptation which are open or constrained (Kirschner and Gerhart 1998; Brown 2014). Some cases of manipulation and deception plausibly rely on evolvability restrictions, for example the orchid-wasp case introduced in chapter 4, where the morphology and perfume of the orchid has evolved to mimic signals which, by being integral to the wasp's sexual reproduction, experience a strong conservative pressure.

These considerations are therefore likely to be significant with respect to signalling strategies, especially in competitive signalling situations – such as predator-prey signalling systems or courtship displays among conspecifics. Signallers and receivers come to not just update their strategies, but to do so at faster or slower rates depending on the nature of the strategic encounter they are entwined in, and the mechanisms which instantiate communicative

---

commitment signalling become a valuable and evolutionarily significant alternative, in terms of the diminishing reliability of reciprocal altruism with increasing population?

behaviours and strategies. Even in David Lewis signalling games (along with games of common interest more generally) the Red Queen effect might have a role to play. First, as argued, the precise cognitive mechanisms and procedures employed by senders and receivers are likely to be different. Different systems (or systems utilized differently) will admit to different degrees of plasticity and evolvability – and will have a different set of cross-cutting tasks and utilities exerting distinct demands and pressures. They will differ in the degree that they are cross-linked with other evolutionarily pressured capacities or rely on traits that are generatively entrenched in development. Quick and easy signalling responses will have different pathways of update and adaptation than the (typically) more complex set of systems which appropriate receiver responses require. This of course is the speculation on the basis of *fakeable* signals – the opposite might be true for index signals, where the signals are more diaphanous indicators of underlying state (emotional or otherwise), and so signalling strategies perhaps less flexible (depending on the constraints that enable it). We should therefore not assume that the evolution of sender and receiver strategies always proceeds at the same pace.

There is at least tentative confirming evidence of asymmetry between sender and receiver roles in the literature on great ape communication. For example, (Hobaiter and Byrne 2014) stress the great sophistication and flexibility on the receiver side of Chimpanzee gestural communication, while (Seyfarth and Cheney 2003) find that greater inferential sophistication on the receiver side is a feature of many primate communication systems. Speculatively, one interpretation of this would be that great ape sender strategies are more direct, reactive and expressive, with responses being more strategic or learning-sensitive in comparison. I.e. great ape communication might skew more toward index signals, with many displays underwritten by hard to fake emotions. This gels with the general impression of the volatility of chimp society, where navigating dominance hierarchies (for all concerned) would be aided by having a more automated indication of who is on the warpath right now. So, while these findings do not directly support the structural and evolutionary responsiveness concerns, they show that real-life sender and receiver strategies (in our near biological cousins at least) exhibit important differences, suggesting cognitive asymmetries in the general spirit of those concerns.

It also seems plausible to expect asymmetry in highly culturally dependent signalling systems, such as those that the RST predicts. Consider for example a high-intensity ‘spectacular’ ritual at a religious festival, such as the possession or piercing rituals described in (Power 2017). The performances of religious rituals are intricately constructed, based on learning and observation

of previous performers, and infused with the individual participant's fervour and interpretation of the ritual. But this is just on the sender side. The responses of the receivers are (by hypothesis) just as much a part of why the ritual evolved, but the receivers are all the other participants caught up in the buzz, noise and spectacle, and their responses are far more reflexive. An uneducated, naïve foreigner dropped into the middle of this religious festival wouldn't have a hope in hell of carrying off a convincing ritual performance. And they will also lack understanding of the context of what they're witnessing, and so may experience alarm in response to performances that locals understand as deeply meaningful. But much of their reaction to what they see will also be in common with the locals. They will experience many of the same reactions of awe and fascination in that intense atmosphere, and the performers will seem impressive, mysterious, and obviously profoundly devoted to whatever it is they are doing. In short, sender strategies are highly dependent on cultural learning, to the point of being impossible without such a mechanism of transmission. But the receiver strategies here seem less fluid and tied (to a greater degree) to more human-general cognitive mechanisms that are (presumably) more deeply entrenched. It is of course possible that these are cognitive biases being maladaptively exploited (for example with the CREDs explanation). But the RST prediction would be that the capacity to experience awe and be impressed (rather than repelled or amused) by extreme displays of devotion evolved in response to such displays, as part of an adaptive response strategy. If there *is* such an evolutionary history to the response, then it evolved on a much longer timeframe than the sender strategies and signalling forms themselves.

This is of course speculative, but if we do want to take the RST seriously then such considerations provide rationale to look at how the asymmetry of senders and receivers might be modelled, with respect to signalling robustness in the face of such de-idealisations. In summary, there are two structural modifications to consider which would be especially salient: the addition of 'hedgehog' strategies for receivers, and differing rates of change in sender and receiver strategies.

7.4. The evolution of signal form

One final question suggested by the earlier chapters on RST concerns signal form. We saw that the choice of fakeable or unfakeable models of signalling can imply significant differences between what RST gets to predict (especially with respect to how signal costs should evolve)

and the plausibility of its potential applications to real-world religious behaviour. What has not been addressed in detail so far is the comparative evolvability of these two basic signal forms over relevant parameter ranges. It might seem, for example, that the availability of a constrained, hard to fake signal would make such a signalling system quite likely, or at least a costly one redundant. Formal modelling, at least in principle, might be able to better inform such speculations. For example, how likely is the evolution of index signals compared to costly signals, assuming similar costs and benefits? If we observe putative religious signals with certain costs attached to them, how likely is it that they are of one form or the other? Answering this could make a real difference for any efforts to model religious signalling in the real world; as we could make reasonable efforts to 'triage' our modelling options.

Some of the existing literature here is instructive. For example, have already seen that the Frank's full disclosure principle rests on a flawed assumption. Even constraining signalling so that false signals are strictly impossible does not guarantee that high type senders will inevitably conform to signalling honestly (on fear of being treated as a low type), as even small efficacy costs make available an evolutionarily significant pooling equilibrium (J. P. Bruner 2015a). What this result depends upon is nature not being perfectly even-handed with respect to the distribution of types, i.e. when high types are more common than low types there can be evolutionarily stable configurations of strategies where receivers simply treat all senders as high types. Another surprising result from this is that some degree of deception can take place as well, if the 2x2x2 assumption is further relaxed and there are more than two types with uneven distributions (in his example, high, low, but also medium) – even if each type's signal is unavailable. Bruner's example here is neat and bears repeating:

> "Receivers acquire a payoff of 1 for correctly identifying the sender, and senders receive a payoff of 1 if they are identified as a high type, 0.5 if identified as a medium type, and 0 if identified as a low type. If 20% of the population are low types, 40% are medium types, and the remaining 40% are high types, the following arrangement is stable. High types send their type-specific signal, while both medium and low types do not send their type-specific signal. Upon not receiving a signal, receivers assume their counterpart is a medium type. This arrangement is an equilibrium. Receivers do best to assume their counterpart is a medium type since medium types outnumber low types. Furthermore, medium types have no incentive to send their type-specific signal." (J. P. Bruner 2015a, 660)

The deception here, with low types piggybacking on a lazy signalling strategy by medium types (which efficacy costs would more greatly incentivise), means that even idealised index signals,

reliable by design, should be seen as options in a marketplace rather than a sure bet. Receivers (especially) could well be in the market for something else, and it is not intuitively clear at what point costly signals (for example) might become competitive, in terms of efficacy vs strategic cost, and realistic degrees of constraint (i.e. fakeability – the chance of a non A-type sender successfully sending an A-type signal). It is even less clear how such considerations would pan out in terms of evolutionarily stable strategies and evolutionary significance.

One study which does compare the evolutionary significance of costless and costly signals is (Simon M. Huttegger, Bruner, and Zollman 2015), which modifies the standard costly signalling game (figure 5-4) to include both a differential cost and a differential rate of signal success. What signal success/fail rate models is fakeability; i.e. this is a way to model index signals, or handicap signals, or mix the two forms arbitrarily by modifying the parameters. More formally, each signal attempt by high (w_1) or low (w_2) types has a signal cost c_1 or c_2 as before, but now both also have a success rate parameter s_1 or s_2, which represents the probability that the signal fails to be observed by the receiver (so it appears as though no signal was sent). We can see that if $s_1 = 1$ and $s_2 = 0$, then this models a costless, *impossible* to fake signalling game (the low type can never send the signal). In this case we can interpret any signal costs c_1 and c_2 as efficacy costs (which may or may not be the same). The standard costly signalling game is equivalent to this game when $s_1 = s_2 = 1$, and intermediate values define signalling games that are a blend of the two basic signalling forms. These parameters therefore allow a generalisation within which costly signalling and index signalling can be seen as special cases.

Huttegger et al use this game (dubbed the Pygmalion game) can to explore the basins of attraction for a range of parameter values, manipulated exogenously. They find that depending on the values of these four parameters and the relative prevalence/probability of high types p, the Pygmalion game exhibits two different types of communicative equilibria. One conforms to the standard costly-signalling separating equilibrium: high types (only) signal as high types, and receivers treat senders as high types only if a successful signal is observed. The other is an index signal-like “pseudo-separating” equilibrium, in which both sender types signal and the receiver treats the sender as a high type only if a successful signal is observed. I.e. all senders try to send the signal, but the high types are more likely to succeed. In the separating equilibrium, differential cost is the stabilising force, in the pseudo-separating equilibrium it is the differential probability of signal success.

As might be expected, Huttegger et al. find that these game equilibria exist over different (but overlapping) parameter spaces. Through simulation, they are also shown to vary in their relative evolutionary significance over those spaces; with the separating equilibrium for example being more robust in the face of escalating signal costs for high-type senders (but at lower values of p). What the results do not show is any dramatic difference in evolutionary significance: in broad terms, both forms of signalling are potentially viable, and neither dominates the other except at extreme parameter values.

However the conclusions that can be drawn here are limited. Because the form-setting parameters in the model are exogenous constraints rather than evolvable traits of the signals and population, it is not the case that different signalling forms are ever in competition with one another – there is just one signalling game with two kinds of equilibria. So, while each parameterisation of the game to some degree 'blends' the structures of costly and index signalling games, it does so into a single 2x2x2 structure, and the sender still only has the choice between the high-type signal and the null signal.

The likelihood of different signalling forms might look very different if sender strategies could choose from moves that included a) no signal, b) costly signal, c) index signal, and so-forth. One signal form would be able to much more clearly dominate the other, so these sort of game structures would be better models of the evolution of signal form in the wild. Ideally, they could be expanded to include other signalling options as well, or options with endogenous variables, such as allowing signal cost to freely evolve generation by generation, as a continuous variable. In principle, building in the endogenous evolution of *cooperation* as well (i.e. by letting the average fitness of high and low type senders, given their relative total expected payoffs down the w_1 and w_2 branches) would allow an even deeper insight into what kind of signalling models (if any) can generate the sort of behaviour that RST needs. Such work hasn't been done yet, but seems tractable even with relatively simple, idealised signalling game methods building off of those described, and would do much to inform the theoretical (and potentially empirical) aspects of religious signalling as an explanation.

7.5. Summary of chapter 7

The goal of this chapter was to demonstrate that there are questions relevant to RST which only formal modelling (and especially computational modelling) can address. This is largely because of the trans-disciplinary nature of RST: positing a role for signalling in the evolution

of religion and cooperation will imply predictions based on the evolutionary significances of plausible signalling models. But some of these predictions are far from obvious. Modelling work can therefore help better flesh out the sort of predictions that different interpretations of RST might end up making, and indeed the more general empirical commitments of the approach. Simulations are therefore like experiments in that their results can supply new information about the fit between model and world (and between model and the original theory-driven intentions for using it). In some cases, formal results may necessitate or allow theoretical revisions that either weaken the theory or allow it to be expanded.

In arguing this, I identified three examples of the sort of questions where formal modelling might inform the contours of the possible with respect to RST. In broad summary, these were:

1. How does the co-evolution of signalling with cooperation affect the overall likelihood of the evolution of religious signalling?
2. Does the sort of sender-receiver asymmetry we might expect in the RST context make the evolution of signalling systems more likely or less?
3. How evolutionarily plausible are fakeable or unfakeable signal forms, in relative terms (and again, over parameter values relevant to the RST context)?

These are just three examples, the selection of which was a function of the literature available – that are not in any way an exhaustive list. But the partial answers that we can extract from the existing literature, for each of these three questions, informs various versions of RST (and the general approach itself) as serious proposals. More complete answers would do so to a much greater extent. But they can only really be addressed via further formal modelling work. In the next chapter I will present the methods and results of some original modelling work which takes a closer look at question 2, the asymmetry question, and draw some further tentative conclusions about the evolution of religious signalling.

8. Simulating religious signalling: results and discussion

In this final substantive chapter, I present the results of two sets of simulations which investigate how normal basins of attraction for several relevant signalling games are altered by the introduction of asymmetries between the evolutionary options and dynamics of senders and receivers. The motivation for this was outlined in the previous chapter (in 7.3). Briefly, there is independent reason to suspect that different mechanisms and strategic situations might reify sender and receiver strategies in religious signalling, and these differences may be evolutionarily significant. Following the motivating reasoning laid out in 7.3, and further interpretive arguments to be discussed, we can make corresponding modifications to simple signalling game models, and test how the basins of attraction for signalling equilibria are impacted by those modifications. Do plausible differences in the way that sender and receiver strategies update mean that signalling is more likely or less? In addition, the final set of results investigates hybrid and pooling equilibria for costly signalling games, and what they might mean for RST-relevant signalling behaviour.

8.1. Methods

Methodologically speaking, these simulations all use approaches also discussed in the previous chapter, including robustness analysis (measuring the proportion of randomised initial populations that reach the relevant signalling equilibrium for a given model under specified parameters). They use customised signalling games based on the standard signalling games laid out in chapter 5, with alterations to match the specific motivation cases. This section specifies and elaborates on the methodological details that are common to all simulations reported in this chapter, with target-specific methods and results reported in 8.2 onward.

8.1.1. Model architecture

Populations, games, and strategies

The simulations all use infinite, two population models. Sender and receiver populations are each represented as by a single vector with one element for each strategy, representing the proportion of that strategy in the population. I.e. if there are four sender strategies, then the sender population is a vector of four numbers summing to 1. The initial proportions of the different strategies within sender and receiver populations are randomly generated.

The signalling games used are all derivations of the 2x2x2 signalling games discussed in chapter 5, defined by the sender-receiver game structure and fixed payoffs, with signal costs and world-branch probability (i.e. probability that nature plays w_1) varied exogenously as parameters.

The fitness of a given strategy at a time period t, for the specified game and parameterisation, is determined by the expected payoff of an interaction with each of the strategies in the opposing population weighted by the probability of encountering that strategy (i.e. the relative frequency of the opposing strategy in the opposing population). Because some strategies conditionalize on the state of the world, the probability of the world being in w_1 or w_2 will influence these fitness values.

Evolutionary models

These payoffs at time t are then treated as finesses and relativised to one another to drive the evolution of strategies within populations. The proportion of each strategy in the nominal next time period $t+1$ is determined by the standard discrete-time replicator dynamics. For the sender population this is:

$$X_i(t+1) = X_i(t)\frac{F_i^S}{\bar{F}^S} \qquad [8.1]$$

where X_I is the proportion of senders using the ith sender strategy at a given point in time, F_i^S is the fitness of that strategy and $\bar{F}^S$ is the average fitness of the sender population. Likewise, for receivers:

$$Y_j(t+1) = Y_j(t)\frac{F_j^R}{\bar{F}^R} \qquad [8.2]$$

where Y_j is proportion of receivers using the jth receiver strategy, F_j^R is the average fitness of that strategy and $\bar{F}^R$ is the average fitness of the receiver population. In order to smooth out discrete jumps, the simulations calculate the vector of change for each strategy, i.e. $X_i(t+1) - X_i(t)$ for senders and $Y_j(t+1) - Y_j(t)$ for receivers, which at each time-step is applied to the strategy proportions using a small increment factor, typically 10^{-2} or less. These vectors are re-calculated, and the steps repeated as the two populations evolve in reaction to each other, until a maximum number of time-steps set to allow populations to settle into stable arrangements. The update process is deterministic, with no randomization or mutation.

Robustness testing

For purposes of robustness testing, at least a thousand of these simulations are run for each parameterisation of interest (i.e. each simulation run only differing with respect to in the initial randomised populations). The proportion of those simulations which reach the available communicative equilibria (separating or hybrid) are tallied. Parameter values are sampled in order to step across the relevant parameter space for the signalling game in question and map out the basins of attraction for signalling. This allows insight into how the likelihood of signalling co-varies with parameter change within a single signalling game; and allows comparison between different games.

8.1.2. Implications and limitations

This methodology has strengths, weaknesses, and implicit assumptions which should be acknowledged and clarified. First and foremost, it is a relatively simple architecture, which is efficient both in terms of coding complexity and computational demands. The trade-off for this is the inclusion of some highly idealised elements, especially with respect to the population model and evolutionary dynamics. However, the implications of these idealisations are not straight-forward.

For example, the population is continuous rather than agent-based, meaning strategy proportions are infinitely divisible. The obvious worry is that this might approximate a large population but is a problematic idealisation in the context of traditionally small-scale human interaction groups. It also means that strategies never truly reach fixation or extinction, only approaching proportion 1 or 0 asymptotically (due to the mathematics of the replicator dynamics, so that tiny, residual levels of each strategy always remain. Importantly though, what the vector of strategy proportions represents within the simulation is just the overall chance of a strategy being encountered, and this can be realised in several ways. For example, the strategy vector S = [0.5, 0.2, 0.3] could equally represent i) an infinite population divided into three strategy-classes of agents, present in those proportions, ii) a homogeneous population of any size (infinite, finite, or just a single agent) where all agents play the same mixed strategy: mixing the three pure strategies with probabilities according to S, or iii) a heterogenous population (finite or infinite) where each agent mixes strategies such that S describes gives the chances of any given strategy being played in a random encounter. The continuous population models represent the prevalence of *strategies*, not agents, and they are agnostic with respect to how real populations of agents might combine to realise those strategy proportions. The results

here are therefore quite general, and strategy fitness would attach to agents in proportion to their playing those strategies. This is a simplifying assumption permitted by the idealised infinite population.

All this is to caution against porting-across ontological assumptions e.g. from population biology, where gene frequencies are the target of enquiry. If we were modelling the evolution of an 'on-or-off' trait, or one carried by a single allele, then the lack of small-numbers behaviour like fixation or extinction might indeed be a concern for fit between model and world. But behavioural traits such as prosociality and signalling strategies are not plausibly like this; they are inherently messy and probabilistic. Even the most behaviourally consistent real-world human agents maintain a capacity to surprise (even if the probability of doing so is low). Idealised as they are, continuous population models divided into strategies rather than agents are appropriate for the target system in question.

Finally, the universal use of the replicator dynamics (as opposed to also testing using learning dynamics) might be questioned. But this part of the evolutionary model too has flexible interpretations. The discrete generation time is most obviously interpreted as corresponding to a biological generation, but this is an over-interpretation – as showed by the variable 'smoothing' function discussed above. Generations here are entirely artefacts of the modelling choice. In a cultural interpretation of the discrete-time replicator dynamics, a generational step is just a time-step at which the population is being sampled – during which strategies have been altered in a way that approximates a selection mechanism. No assumptions are made about the actual mechanisms which reify the selection-like evolutionary dynamics of the population. As reported in (Skyrms 2010), evolutionary signalling models tend to be robust under variation of formal evolutionary dynamics (i.e. results using replicator dynamics can be reproduced using other dynamics), so the type used will remain fixed here, as a simplifying assumption.

8.2. Sender-receiver asymmetry

The first series of simulations address the two concerns raised about sender-receiver asymmetry from 7.3. This will be done via two modifications to standard signalling games: i) the addition of a 'hedgehog' strategy for receivers, and ii) varying the relative speed of evolution between sender strategies and receiver strategies.

8.2.1. Lewis model and interpretation

As discussed in 7.3.3, much of the literature on the effects of de-idealisation on signalling has centred around costless signalling and coordination games, modelled using the David Lewis signalling game. Because of this and its relative simplicity, this is also the first signalling game model I will investigate for asymmetry effects. The extended form of the game was given in figure 5-3: it is a 2x2x2 game with no signal costs, and payoffs are normalised to 0 and 1 in a structure which delivers the same payoff to both players. In this game, the two states of the world w_1 and w_2 represent 'neutral' world states, rather than high or low sender types. They can be interpreted as sender-independent world-states, such as in the epistemic-pragmatic mismatch example given chapter 7, where one of a pair of foragers can see the prey animal, and therefore has the information that the other needs to deliver the payoff for both. But world-states in this model could also be states of the *sender*, which the receiver can 'deal' equally well with as long as an appropriate action a_1 or a_2 is performed; and sender and receiver both desire the same action-state combination.

Consider a trivial example: the receiver is buying a pizza for both agents to share and is indifferent whether to get peperoni or vegetarian. But the receiver owes the sender lunch, knows the sender will have a preference (though not which preference), and strongly prefers to order according to the sender's preference. In this case the world-state is the sender's preference and matching the receiver's action to that preference is preferred by both agents – hence the need for some sort of signal from the sender. If the only available signals are arbitrary with no pre-established meaning (assume that the communication takes place in the form of gestures across a noisy/crowded pizza joint, and the sender is terrible at charades), we can imagine a private signalling convention emerging via trial-and-error over many iterations of this scenario.

Other interpretations more relevant to the cooperation problem might involve cooperation-impacting *normative compatibility* between sender and receiver. As a general schema for this, we can imagine a task that would be more productive to carry out jointly as long as the sender's way of doing it was compatible with the receiver's, otherwise it would be better for both if they carried on individually (or found different partners). If the receiver committing to the joint strategy obliged the sender as well (or made repairing to a solo-working option more difficult), then we have a target system that fits the Lewis model. Something like this can be seen in various collective action problems (it is most efficient for me to wash and you to dry, but only

if you put the dishes away where I can find them, etc). It would be reasonable to imagine the emergence of some sort of correlation between the cooperation-impacting practices and otherwise arbitrary, more observable signal traits (if such scenarios were common enough).

Therefore, though the rationale for RST in this thesis is as a potential solution to ancestral (prisoner's dilemma) cooperation problems, there is also the potential to consider a more general role for 'norm-signalling' in solving coordination problems as well, especially if there are ways these contexts might become entangled. I will return to this idea in the discussion.

For the moment though, we can characterise the 2x2x2 Lewis signalling game as involving two world states (w_1 and w_2), a world-observing sender with two possible signals (m_1 and m_2), and a signal-observing receiver with two possible actions (a_1 and a_2). If the receiver's action matches the state of the world, then both senders and receiver get a fixed positive success payoff, otherwise their payoff is zero. Note that the two options for both sender and receiver might correspond to two distinct 'active' options, or a single active option and a passive move (i.e. doing nothing), there just needs to be a difference in observability and payoff outcome, respectively.

Senders and receivers each have the four pure strategies for 2x2x2 signalling games available to them, as discussed in chapter 5, and listed here in table 8-1.

Table 8-1: List of pure strategies for 2x2x2 signalling games

Strategy label	*Strategy description*
S1	Send signal m_1 if w_1 and send signal m_2 if w_2
S2	Send signal m_2 if w_1 and send signal m_1 if w_2
S3	Send signal m_1 unconditionally
S4	Send signal m_2 unconditionally
R1	Perform action a_1 if m_1 and perform action a_2 if m_2
R2	Perform action a_2 if m_1 and perform action a_1 if m_2
R3	Perform action a_1 unconditionally
R4	Perform action a_2 unconditionally

The two signalling equilibria of this game are S1-R1 and S2-R2: these are Nash equilibria where neither sender nor receiver have incentive to defect. Note that this is also the case for four pooling equilibria combining unconditional strategies S3 or S4 with R3 or R4: there is no

incentive to change to conditional strategies because they will not be recognised. However, this is just in the static game analysis. In an evolutionary model with the replicator dynamics operating on continuous populations (and no world-state probability bias), even tiny proportions of S1 or S2 strategies in the sender population will incentivise a growth in the matching R1 or R2 strategies (and vice versa) one of S1-R1 or S2-R2 is inevitable (Simon M. Huttegger 2007).

This will be the case at least as long as the world-state probabilities for w_1 and w_2 are even. As discussed in chapter 7, highly biased world-state probabilities make the evolution of signalling systems less likely. This has been demonstrated for the David Lewis signalling game, including in (S. M. Huttegger and Zollman 2013), where the basin of attraction for the S1-R1 and S2-R2 signalling systems is shown to drop off either side of $p = 0.5$, in upside-down u-shape (though even when p is very close to 0 or 1 the number of populations reaching signalling equilibria is still large). The populations which fail to reach signalling equilibria are those where the initial proportions of matching conditional strategies (e.g. S1 and R1) are low enough that it is preferable to switch to the unconditional receiver strategy R3 or R4 that does best given bias towards w_1 or w_2. The incentive for senders to adopt a conditional strategy then disappears and the community is locked into a pooling equilibrium with unconditional strategies dominating the population.

Three modifications are made to this base model. The first modification is to add a hedgehog action a_3 for the receiver, an unconditional strategy with a variable payoff that is not sensitive to state of the world. As per 7.3, this will test Sterelny's suspicion that signalling will suffer if the strategic situation is made more realistic by giving the receiver something else to do other than listen to the sender. The second is to vary the rates of generational change of sender and receiver strategies relative to each other, via independent modification of the increment factor in their replicator dynamics. This will test the possible effect of having reifying mechanisms for sender and receiver strategies with different evolutionary timescales and update rates. Finally, the bias of nature is also varied (as just discussed), and the effects of these three departures from the David Lewis 2x2x2 game are analysed, with respect to the impact they have on the evolutionary significance of signalling systems.

8.2.2. Hedgehog strategy

Turning the first modification, the receiver now has three possible actions upon observing the signal: a_1, a_2, and now a_3. As before, a success payoff of 1 is received by both players in the case that the receiver plays a_1 while the world is in state w_1, or the receiver plays, a_2 while the world is in state w_2. A payoff of zero is received if a_1 or a_2 are played otherwise. A payoff of h is received unconditionally if the receiver plays a_3, where the value of h is set exogenously between 0 and 1. This allows the receiver a fifth strategy R5, which is "Perform action a_3 unconditionally". In principle, the introduction of action a_3 permits many further conditional strategies, however these are omitted for sake of simplicity[72]. The sender retains her four familiar strategies.

This modification can be given an illustrative interpretation via the pizza-ordering toy example. The sender knows their pizza preference (w_1 or w_2) and remotely gestures (m_1 or m_2). The receiver can't read minds but observes the signal and can choose to order a pepperoni or vegetarian pizza (a_1 or a_2, getting the preference right or wrong), or alternatively play it safe and just order starters (i.e. the hedgehog strategy of disregarding the signals and performing a_3 unconditionally). Neither player prefers the starters to the preferred pizza, but they prefer it to the dis-preferred pizza. Varying the prior probability of the world is equivalent to it being more or less likely that the sender is (for example) an observing vegetarian or die-hard carnivore.

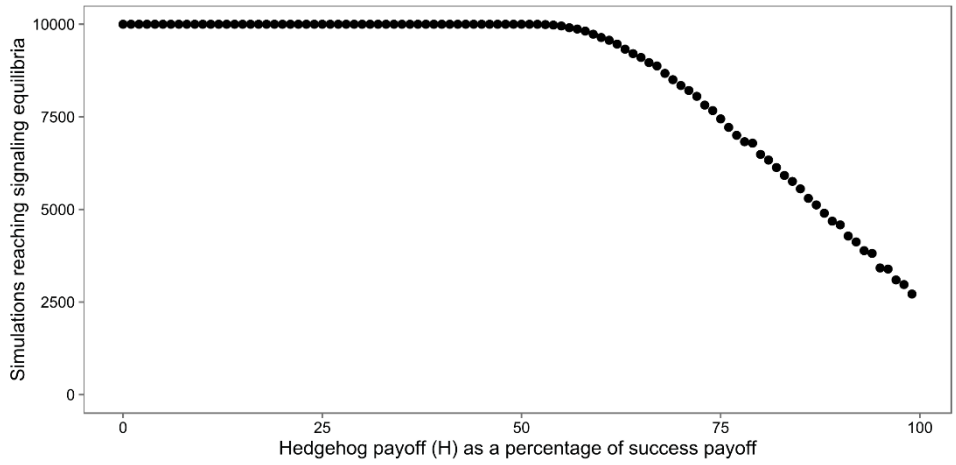
Results

As just discussed, in the simple 2x2x2 signalling game with no world-bias, one of the two signalling equilibria is guaranteed to be reached under the replicator dynamics. Increasing the bias of the world (i.e. making w_1 more probable than w_2 or vice versa) will undermine this, with an increasing proportion of populations instead collapsing to pooling equilibrium. Not surprisingly, a similar effect was found with the hedgehog strategy as values of h, the payoff for a_3, become significant. The hedgehog strategy provides the receiver with an additional unilateral response that can attract some proportion of initial populations away from signalling equilibria strategies when h is in excess of 0.5 (i.e., the average payoff for 'guessing' in the 2-action game). This result (for an unbiased world) is illustrated in Figure 8-1, showing the

[72] These conditional strategies will also strictly be dominated by strategies R1 to R5 in the parameter ranges of interest. For example, the conditional strategy "Perform action a_1 if m_1, and perform action a_3 if m_2" will always give a lower payoff than R1, as long as h is less than 1.

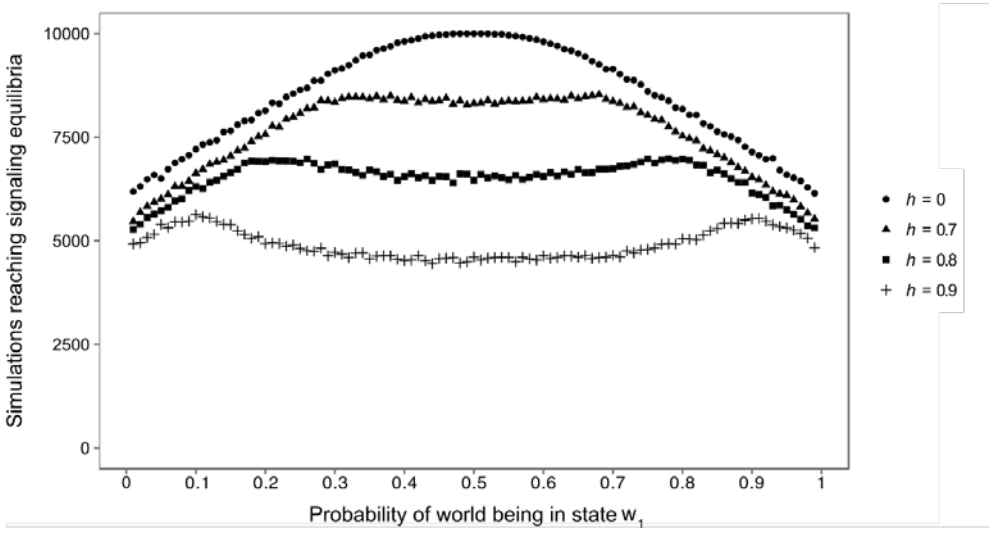
number of simulations out of 10,000 which reached signalling systems for each value of *h* from 0.01 to 0.99 in increments of 0.01. Note that the exact parameter range of this effect, including the point at which the effect becomes significant and the y-intercept, are artefacts of the number of world-states and strategies in the model and therefore not fully general.

FIGURE 8-1: Robustness of signalling with varying hedgehog payoff in the David Lewis game



A more surprising result is observed when world-bias $p = P|w_1|$ and *h* are varied in combination. Figure 8-2 shows the results of varying *p* from 0.01 to 0.99 for selected values of *h*. As expected, at $h = 0$ signalling is guaranteed when nature is unbiased, but progressively less likely to emerge as bias increases, showing a flattened, inverted 'u'-shaped curve. For higher values of *h*, however, we see flattened 'm'-shaped curves instead. Increasing bias does not immediately reduce the likelihood of signalling. In fact, we observe a plateau in signalling behaviour followed by an *increase*, peaking where bias is equal to the value of *h*, i.e. at $p = h$ and $p = 1 - h$. Inside these peaks, all populations either reach a signalling equilibrium or a Hedgehog equilibrium (in which all receivers have adopted the hedgehog strategy). For values of *h* outside of the peaks, where the drop-off commences, the non-signalling equilibria are traditional pooling equilibria instead. At the $p = h$ and $p = 1 - h$ points themselves, half of the non-signalling equilibria are traditional pooling and half are hedgehog pooling.

Figure 8-2: Signalling robustness across world bias at four levels of hedgehog payoff



To see why we get this result, it is worth dwelling on how signalling strategies evolve in signalling games like this. The key point to note is that it is the *receiver's* action which secures the payoff (for both players) in the context of the background world-state probability. This interaction with world-state introduces a basic asymmetry between sender-receiver with respect to the significance of unconditional strategies. For example, if p is high then R3 is an intrinsically valuable strategy, however it is never the best move for the sender to switch to S3, since unconditionally signalling m_1 provides no information, and can only 'accidentally' prod receivers in the right direction if R1 was implausibly high[73].

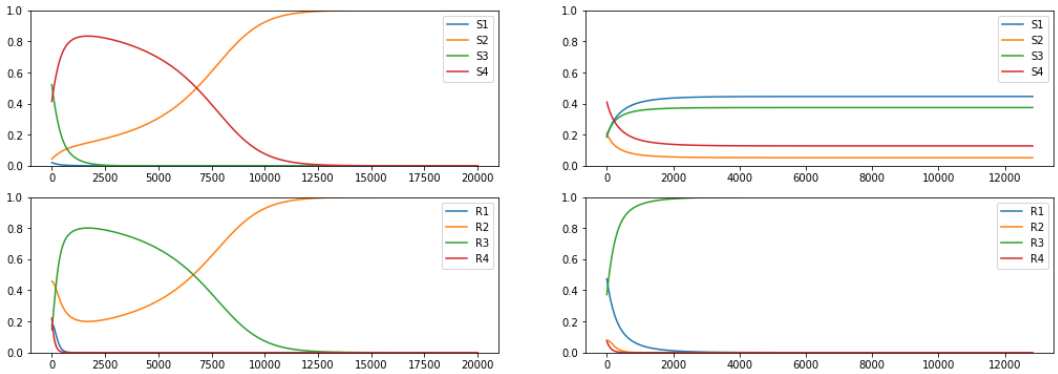
This means that senders' and receivers' pathways for improvement are different, and their strategies evolve in different ways. The sender improves the payoff by becoming more clearly conditional, with respect to the signals that receiver behaviour is most sensitive to (i.e. with respect to the prevalence of R1 and R2 conditional strategies). The receiver can improve the payoff either by becoming more conditional in response to the sender's conditional strategies, *or* by better guessing the state of the world. Whether a signalling equilibrium is established depends on this race for dominance between the receiver's two options, and that in turn depends on the prevalence of conditional strategies in the sender population to incentivise the first

[73] And even then, it could only be less or equal to the fitness of S1. By term 'intrinsic' here, I mean an advantage that is not merely fleeting or washed out in the averaging process of many random populations. All things being equal, it is better for a sender to become more communicative rather than less.

option. Once one conditional sender strategy grows to a high enough proportion, the fitness of the matching conditional receiver strategy (which rises toward a maximum of 1) will overtake the (fixed, less than 1) fitness of the fittest pooling strategy, and the population will inevitably collapse to a signalling equilibrium.

Figure 8-3 shows two representative simulation runs from the David Lewis signalling game, illustrating the evolution of signalling (left) and pooling (right) strategies out of two randomised populations in a highly biased world (p = 0.9). Time runs from left to right, with strategy proportions starting out at random, with sender strategy proportions displayed above receivers. In the signalling population, the S2-R2 signalling system evolves despite the initial fitness advantage of R3 (which is reversed by the growth of S2). In the pooling case, the initially high R1 incentivises S1, but R1 is supressed before the favour can be returned, and so the S1-R1 signalling system never establishes. Note that in pooling cases the evolution of sender strategies ceases once the genuinely receptive strategies R1 and R2 have been suppressed (as there is no further incentive to improve). This 'abandonment' of the senders is characteristic of pooling in signalling games, and demonstrates the dynamical asymmetry of senders and receivers, as receiver strategies do not become irrelevant in this way.

Figure 8-3: Two simulation runs from the David Lewis signalling game (p = 0.9) which result in a S2-R2 signalling equilibrium (left) and a pooling equilibrium (right)



One upshot of this asymmetry is that giving the senders more time to 'get their act together' should increase the likelihood of signalling (on average, all else being equal). One manifestation of this becomes visible merely due to differing numbers of strategies. Given the replicator dynamics, the growth rate of a strategy depends on relative fitnesses but also its initial proportion, and with five rather than four receiver strategies each starts off (on average) 20% lower in proportion. This means that receiver pooling strategies will have less time to

quash rival conditional strategies before the matching sender strategies adapt to them, so even the $h = 0$ curve 'drops off' more shallowly than it would with only 4 receiver strategies in play.

The hedgehog variant game adds further idiosyncrasies. In particular, the way that the pooling equilibria exclude each other (except at the peak points) is easily explained by which of p, $1-p$, or h is the greater, because these values determine the payoffs for strategies R3, R4 and R5 respectively (information from the sender is irrelevant in the competition between them). Understanding this (in combination with the race between senders and receivers) helps make sense of the overall shape of the curves. First, when one pooling strategy dominates the others, it does so from the very start of the simulation, while (on average) signalling strategies take time to become competitive (from an initial average expected fitness of 0.5, and then senders adapt to whatever conditional strategies exist in the receiver population). The advantage enjoyed by the fittest pooling strategy depends in part on how much it leads the other two pooling strategies: the more of a lead it has, the higher its fitness will tend to be relative to the *average* fitness of the receiver population – and this helps determine its initial growth rate. This means that while a second-placed pooling strategy can never win, it being a *close* second can hold back the rate at which the fittest pooling strategy grows relative to other receiver strategies. Conditional receiver strategies can therefore persist for longer, making an eventual collapse to signalling somewhat more likely. This 'spoiler' effect between pooling strategies accounts for the modest peaks in signalling where the fitness of the hedgehog and world-sensitive pooling strategies are the closest.

Because the cause of these results has to do with a race between the evolution of conditional and unconditional strategies – including between sender conditional strategies and receiver unconditional strategies – it makes predictions regarding modifying evolutionary rates. Specifically, it suggests that de-idealising the model by speeding up the rate of sender strategy evolution relative to receiver strategies should make signalling *more* robust rather than less.

8.2.3. Generational asymmetry

The next modification of David Lewis signalling framework is therefore to introduce 'generational' asymmetry regarding the rate at which sender and receiver strategies evolve. This is achieved by introducing "slow-down factors" z_S and z_R to the replicator dynamics in order to control the rate at which sender and receiver populations change over time. The

modified replicator dynamics for sender and receiver populations are represented by the following equations:

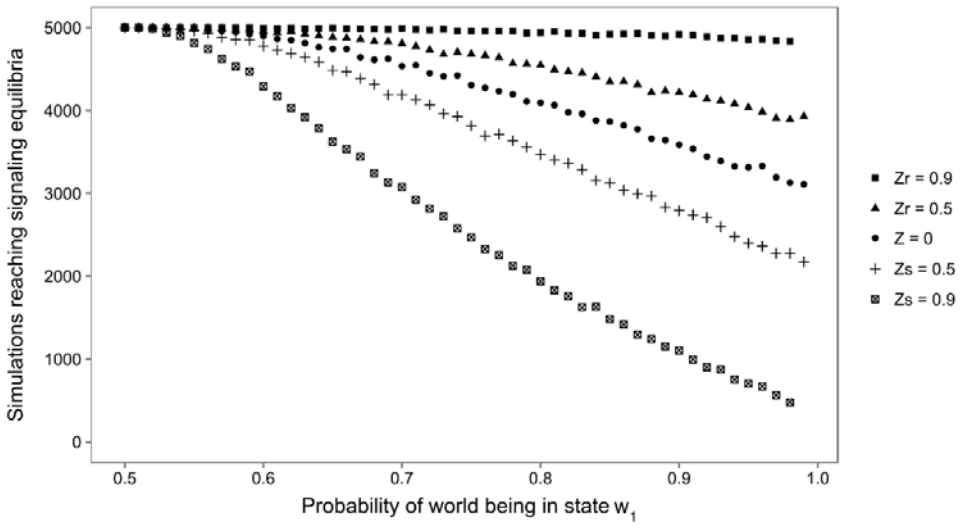
$$X_i(t+1) = (1 - z_S)X_i(t)\frac{F_i^S}{\bar{F}^S} + z_S X_i \quad [8.3]$$

$$Y_j(t+1) = (1 - z_R)Y_j(t)\frac{F_j^R}{\bar{F}^R} + z_R Y_j \quad [8.4]$$

Where z_S and z_R take values between 0 and 1. The right-hand sides of these equations mix the right-hand sides of their regular counterparts [8.1] and [8.2] with the unchanged (current time-step) value for strategy *i* or *j*, in a ratio determined by the slow-down factors. When both z_S = 0 and z_R = 0, we recover [8.1] and [8.2], but the rate of evolutionary change is reduced as they are increased. E.g. setting z_S = 0.5 halves the rate of change for sender strategies, while setting z_R = 1 means the composition of the receiver population remains fixed and only the sender population evolves (assuming $z_S \neq 0$). Again, these equations were implemented in the simulation using a method which provided a smoother incremental change.

Figure 8-4 shows the results of simulation runs where, over a range of world-bias (from p = 0.5 to p = 0.95), where the update rates for either sender or receiver are slowed to one-half and one-tenth, via manipulation of either z_S or z_R. Note that the middle curve here (where $z_S = z_R$ = 0) reproduces the right-hand side of the h = 0 curve from figure 8.2.

Figure 8-4: Signalling robustness across world bias at five relative update rates



As expected, introducing this generational asymmetry between senders and receivers has the effect of directly altering the robustness of signalling, with signalling more likely when sender strategies evolve faster than receiver strategies. Slowing the evolution of the sender population leads to more pooling because, as before, receivers facing a sender population consisting of few conditional senders will do best to simply perform that act that is best suited for the more likely state of the world. The relative advantage of the unconditional R3 strategy means that sender conditional strategies will often be in a race against time to adapt before conditional receiver strategies are reduced to negligible levels and the incentive to improve disappears. This means the relative rate of evolution makes a crucial difference across many simulations with random populations. As figure 8.4 illustrates, this effect is more pronounced for high levels of bias. Slowing the evolution of the receiver population has the opposite effect. Senders now have time to adopt the best separating strategy given the initial mix of receiver strategies. Once this occurs, the receiver population (slowly) adapts and the result is a robust signalling system.

What this also illustrates is that 'damage' done to signalling robustness by attractive pooling strategies can be almost entirely mitigated by setting z_R arbitrarily high, i.e. by letting senders 'take the lead'. Additional exploratory simulation work shows this to also be true for the de-idealisation effects of the hedgehog strategies seen in 8.2.2 (as should be expected).

The implications of this are intriguing. Breaking the symmetry between senders and receivers often significantly reduces the likelihood that a separating equilibrium emerges. In the previous set of simulations, providing receivers with a safe third option allowed a decent payoff to be secured (regardless of the state of the world), and this significantly reduced the size of the basin of attraction of the separating equilibrium. Likewise, separating is a remote possibility when receivers rapidly outpace senders in the race to adapt, in the presence of viable non-signalling strategies. But breaking symmetry in the opposite direction has the opposite result.

8.2.4. Asymmetry in handicap and vulnerability signalling

The next question is whether sender-favouring asymmetry effects can also be found outside the David Lewis signalling game. In chapter 7, I argued that sender-receiver asymmetry should be expected – rapid evolutionary processes such as social learning and success imitation might play a far greater role in sender-favouring asymmetry effect in evolution of sender strategies. And in chapter 6 I outlined a potential application template for the costless signalling model to

RST. But this was a minority among the identified signalling model templates and was a tentative extension of usual paradigm: RST was introduced in the context of the traditional cooperation problem, with the assumption that the payoff structure of its signalling games will reflect the prisoner's dilemma. It is therefore unclear how significant these asymmetry results from the Lewis game are to RST in general.

In the next set of simulation then, the generational asymmetry simulations of 8.2.5 were reworked using partial conflict of interest signalling models instead. Specifically, differential intrinsic cost signalling and differential vulnerability cost signalling models. The extended forms of these signalling games were given in figures 5-4 and 5-6, respectively. As 2x2x2 signalling games, the available strategies are the same as in table 8-1. The methods are the same: the introduction of generation asymmetry via the use of slow-down factors z_S and z_R in the modified replicator dynamics of equations 8.3 and 8.4. A range of parameter values for p and c_2 (signal cost for low types) were investigated, with c_1 (signal cost for high types) fixed at zero, for simplicity.

The results here are negative and will be summarised verbally. The differences that varying z_S and z_R made to the likelihood of signalling were negligible. Where small differences in signalling robustness were initially found at various parameter ranges, on investigation they turned out to be artefactual and not representative of the model[74].

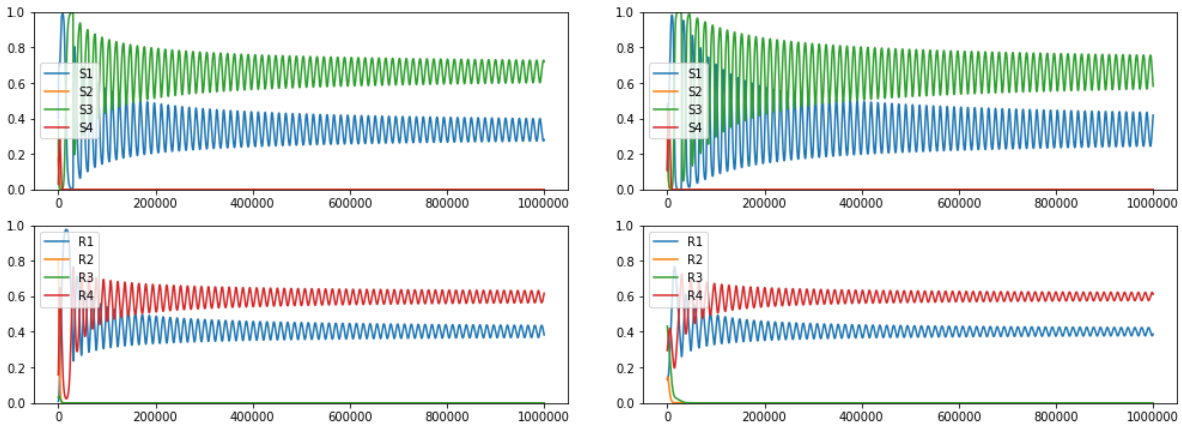
The explanation for this is straight-forward. It is no longer in the interest of receivers to let senders 'take the lead'. Whereas in the Lewis game receivers could count on senders to look after the interest of both agents, in partial conflict of interest models the senders have their own agendas. Increasing the update rate of sender strategies increases both the speed at which high-type senders can coordinate with receivers, and the speed at which low-type senders can exploit that coordination. The consequence of this differs from population to population (depending on initial distribution of strategies), but largely average out with repetition.

---

[74] For example, at extreme values of z_S and z_R there was a small a decrease in the number of populations which reached signalling equilibria. However, this could be accounted for either by populations which evolved too slowly and failed to reach a stable state before the maximum number of generations, or by rapid 'jumps' in the early steps of the simulation which pushed ultimately unstable strategy proportions into values extremely close to one, delaying collapse to final state. When simulation run-time to completion was increased, and/or the increments were made smoother, these artefacts largely disappeared.

This null result holds for hybrid equilibria as well as separating equilibria. In hybrid equilibria, senders mix strategies S1 (conditional signal high if high) and S3 (always signal high), while receivers mix R1 (conditional treat as high if signal high) with R4 (never treat as high). At equilibrium, the mixing rates oscillate in response to one another, in closed orbits around a quasi-stable point in phase space. One possible hypothesis was that asymmetric update rates would alter the informational quality of the hybrid equilibrium; by shifting the stable point or otherwise causing higher or lower average rates of informative conditional strategies. But this does not happen. Figure 8-5 shows the evolution of signalling strategies in the differential intrinsic cost game for parameter values c_2 = 0.4 and p = 0.4, in two different random populations where: a) senders and receivers evolve at the same speed (left hand side), and b) where senders evolve at five times the rate of receivers (right)[75]. Both show the evolution of random populations toward dynamically stable hybrid equilibria, with conditional and unconditional strategies oscillating around a fixed point. Once the oscillations settle into dynamic stability, the only difference between them is in the amplitude of the oscillations. The average proportion of signalling strategies S1 and R1 over time (and hence the quantity of information being transferred) is the same, determined by the parameters (see 5.2.1).

Figure 8-5: Two simulation runs in the differential intrinsic-cost game, showing the effect of increased sender update rate relative to receiver (right) on hybrid equilibria



[75] With the parameterisations used here, the effective update rate for senders in the right hand side is $\sqrt{5}$ times its rate in on the left hand side, whereas the receiver update rate is $\frac{1}{\sqrt{5}}$ its rate on the left.

Therefore, in the differential cost signalling game (traditionally used for modelling signalling with partial conflict of interest), it does not matter if senders or receivers update their strategies faster or slower. While the results here are hardly decisive, they illustrate a few general points. For example, they count against the notion that receivers will be more easily exploited and vulnerable to manipulation *in the long run* if they are less evolutionarily flexible than senders. Manipulation in this dyadic model is an out-of-equilibrium phenomenon, unless $z_R = 1$, i.e. unless something is preventing receivers from adapting entirely. As discussed in the wasp-orchid case back in chapter 4, one avenue for manipulation is if the signalling environment is *not* dyadic as far as the receiver is concerned, and the annoyance of false-positives like a seductive orchid is outweighed by the real cost of not being sensitive enough to genuine, strongly fitness-enhancing signals.

8.2.5. The prevalence of costly signals without separating equilibria

One secondary question we can address in passing is: how significant are non-separating equilibria in driving costly signalling behaviour? The phenomena that signalling theories of religion primarily seek to explain is the prevalence of religion, seen as a signal. However, the prevalence of a signal in a population (in this case sender move m_1) is not necessarily the same as the prevalence of the signalling system which supports it. Nor is it equivalent to quantity of information transmitted within that signalling system (as hybrid equilibria show, the prevalence of communicative signal and communicative response is often highly unequal). This provides licence to look at the simulation results here in an alternative manner.

The robustness analyses up to now has been looking at the number of separating equilibria per parameter point. The work on hybrid equilibria discussed in chapter 5 shows that hybrids also can stabilise levels of costly signalling in populations, and in parameter ranges where separating equilibria are not available (for example, with the value of c_2 being less than that payoff for success). We saw that sender-receiver asymmetry does not increase the level of communicative strategies, but what level of signalling behaviours do hybrid equilibria support and maintain?

For any given state of a population in our simple 2x2x2 models, the proportion of senders which are sending the signal can be calculated as the prevalence of S1 multiplied by the probability of w_1 plus the prevalence of S3 (note that S2 is never a stable strategy). Table 8-2 shows the average values of this calculation for 1000 populations at simulation end-state (after

50,000 generations), across nine parameterisations of world-state probability and signal cost for low types[76]. Notably, the parameter space here is all below $c_2 = 1$, the threshold for classical, communicative separating equilibria where low types are rationally excluded from signalling. The values reported have no meaning within a single population, but rather represent the average expected prevalence of costly signalling at equilibrium. That is to say, assuming a pure strategies interpretation (see 8.1), they correspond to the average probability that a random sender from a random population will be sending the costly signal at equilibrium (regardless of whether that is a communicative or a pooling equilibrium).

Table 8-2: Proportion of sender populations who signal m_1 in the differential intrinsic-cost signalling game, across world-bias ($p = P|w_1|$) and sub-exclusionary values of c_2

	$p = 0.1$	$p = 0.2$	$p = 0.3$	$p = 0.4$	$p = 0.5$	$p = 0.6$	$p = 0.7$	$p = 0.8$	$p = 0.9$
$c_2 = 0.1$	0.20	0.40	0.59	0.80	0.68	0.54	0.51	0.48	0.45
$c_2 = 0.2$	0.20	0.39	0.61	0.80	0.74	0.59	0.51	0.52	0.46
$c_2 = 0.3$	0.20	0.39	0.61	0.79	0.78	0.61	0.52	0.52	0.45
$c_2 = 0.4$	0.20	0.40	0.60	0.79	0.83	0.64	0.55	0.52	0.50
$c_2 = 0.5$	0.19	0.40	0.60	0.80	0.85	0.66	0.54	0.49	0.50
$c_2 = 0.6$	0.20	0.39	0.60	0.80	0.86	0.65	0.55	0.53	0.48
$c_2 = 0.7$	0.19	0.40	0.59	0.79	0.88	0.66	0.58	0.53	0.49
$c_2 = 0.8$	0.19	0.39	0.59	0.79	0.90	0.66	0.56	0.51	0.49
$c_2 = 0.9$	0.19	0.38	0.59	0.80	0.89	0.69	0.56	0.52	0.50

Recall from chapter 5 the analytic result that the hybrid (in both differential intrinsic and vulnerability costly signalling games) exists only when the w_1 probability p (of being a high type) is less than 0.5. This means that every population in this range settles into a hybrid signalling equilibrium, with no hybrids present at $p \geq 0.5$. The prevalence of the costly signal action m_1 at $p < 0.5$ where the hybrid *is* available is between 0.2 and 0.8, indicating that the hybrid equilibrium is significant with respect to explaining costly signalling (when high types are less common than not). However, we still see a high degree of the m_1 signal, even at $p \geq 0.5$ when *no* signalling systems (separating or hybrid) are available: all equilibria in this parameter range are pooling equilibria. Indeed, comparing the $p = 0.4$ (hybrid) and $p = 0.5$ (pooling) columns, we can see that the prevalence of the costly signal can be up to 0.9 of the

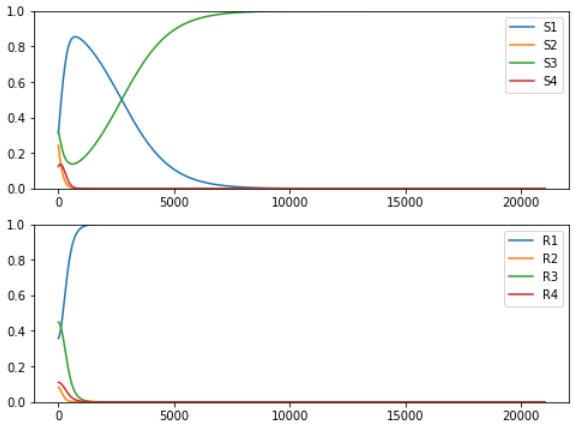
[76] Note that the data here is not very high resolution, and small differences of 0.01 – 0.02 are generally attributable to noise rather than parameter-driven variation.

signalling population when only pooling equilibria are available. The likelihood of the 'signal' (absent signalling) peaks in the middle of the p parameter range, with the signal cost to low types only making a significant difference

The reason we see the costly signal being sent in pooling equilibria this is illustrated in figure 8-6, which shows a simulation run for the differential intrinsic cost game where $p = 0.5$ and $c_2 = 0.9$. There is no separating equilibrium available (as c_2 is less than the success payoff), and this result is a pooling equilibrium. It involves a high initial level of S1 and R1 strategies, which lead to a high level of the R1 strategy for receivers, which is then exploited by low types who also adopt the signal in order to be treated as high types (i.e. S3 dominating the sender strategies). However, receivers go along with this and treat all signalling senders (i.e. all senders) as high types. They do not counter by switching from R1 to R4 (never treat as high), because at $p = 0.5$ it does not improve their payoff, and at $p > 0.5$ it does worse. And they do not switch to R3 even at $p > 0.5$ because doing so would make no difference to them[77], given all senders are already signalling m_1. So, high levels of the conditional R1 strategy force low type senders to signal, making S3 their equilibrium strategy. All populations at $p \geq 0.5$ settled into either this S3-R1 equilibrium, or an S4-based pooling equilibrium where the costly signal is unconditionally avoided (this means that the figures in table 8-2 at $p \geq 0.5$ straightforwardly represent the prevalence of S3 in the sender population). This high prevalence of the m_1 signal persists across the parameter range (though to a lesser degree) when a non-zero signal cost for high types is introduced as well.

[77] It would make a great difference to senders, because then they could ditch the costly signal without being punished for it. But there's nothing in this for receivers so they remain at the discriminating R1 despite any difference in sender moves to discriminate between.

Figure 8-6: **Single simulation run showing the evolution of 'costly lying' pooling equilibrium in the differential intrinsic-cost signalling game (p = 0.5, c_2 = 0.9)**



This S3 'costly lying' pooling equilibrium is a fitness trap for low-type senders that traps the population in a sub-optimal local maximum on the fitness landscape. If the senders could choose between a S3 and S4 equilibrium directly, they would choose S4 hands down. But they don't get to make that decision. Likewise, it would be better for them to not be playing a signalling game here at all. In the case of the figure 8-6 senders, low types would do five times better in terms of expected payoff (0.5 vs 0.1) if signalling had never been an option and receivers proceeded to discriminate by guesswork alone. The initial transient burst of signal-mediated collusion between receivers and high type senders, together with the fact that receivers are indifferent to changing once the informational value of the signal is lost, has locked them into the meaningless lie of the costly signal.

In this way, a costly behaviour might be explained as the pernicious consequence of signals (in the loose sense of the term) and signalling strategies, even when signalling systems themselves are not stable and no information ends up being transferred. The prevalence of this fitness trap is a more pessimistic analogue of the results found in (Skyrms 2002), where costless signals catalysed convergence on the optimal favourable stag hunt equilibrium despite their informational utility disappearing as a single signal was converged upon. In this costly signalling case, the transitory information of early, out-of-equilibrium signalling behaviour leaves a costly legacy for the population. Simply having costly signalling strategies available is enough to drive many populations to costly behavioural conformity, despite the ultimate uselessness of that behaviour.

Between hybrid signalling systems and the fitness-trap pooling equilibria, the rate of costly 'signal-like' behaviour in the simple handicap signalling game is therefore surprisingly high.

This is easily overlooked when traditional, separating equilibria are the only subject of robustness analysis, and we do not look beyond robustness analysis to investigate the out-of-equilibrium dynamics of populations before equilibrium is reached.

8.3. Discussion

There are several points arising from this chapter, which should be summarised and put into overall context.

8.3.1. Hedgehog strategy

First, the presence of a hedgehog strategy predictably makes signalling equilibria less robust in the David Lewis signalling game[78]. The first upshot of this is a vindication of the worries voiced in (Sterelny 2012b); namely that 'walk away options' for receivers should be considered when applying signalling models to real-world target systems. This will be a consideration whenever an audience is not guaranteed. In the case of RST, 'walking away' might include i) simply not paying attention to a sender's religious signalling, ii) opting out of observing religious festivals (if one's presence or otherwise would not be noted or rewarded/punished – i.e. if the receiver did not also have to be a sender), or iii) opting to cooperatively partner instead on the basis of alternative information. As long as the profitability of such courses of action is comparable to that of signalling strategies, they have the potential to stymie the evolution of signalling systems. What will make the difference here is the difference in payoffs on offer. While no firm lessons can be drawn regarding how this would affect the plausibility of RST, it adds yet another contextual factor to bear in mind, and other reason for caution in translating idealised modelling results into anthropological application.

Second though, the results also showed that the signal-degrading effects of de-idealisations are not always additive and can combine and compete in unexpected ways. If world-bias and the hedgehog strategy were additive in their effects, we would see concentric u-shaped curves and much lower levels of signalling when world-bias was high. While it would be wrong to see them as cancelling each other out (at least in this case), the overall effect is less than a sum of its respective parts. This suggests that pessimism about the prospects of signalling should not be additive either. We cannot simply take isolated understandings of the signalling confounds

[78] While it was not reported formally in the text, signalling robustness is similarly degraded by hedgehog strategies in costly signalling games.

we might expect to be present in a given target system and infer an overall estimate of how signalling will be confounded. In principle, this is good news for complex signalling explanations (at least to some degree). In practice of course, it makes realistic estimates more complicated. The lesson by now should be a familiar one: a potential improvement in explanatory potential at the expense of analytic ease and tractability.

8.3.2. Sender-led asymmetry

The results of the generational asymmetry simulations showed that the robustness of signalling results in the David Lewis signalling game can be dramatically altered by breaking the idealised symmetry between sender and receiver evolutionary update processes. In particular, signalling is more likely if senders can evolve at a more rapid rate than receivers. Something like this is arguably seen in other formal literature. For instance, (Hofbauer and Huttegger 2008) find that under the selection-mutation dynamics signalling conventions are possible when the mutation rate of the receiver population exceeds the mutation rate of the sender population. Large mutation rates interfere with the evolutionary responsiveness of a population by introducing deviations away from the optimally adaptive evolutionary trajectory, so one way of interpreting their results is complimentary: signalling is more likely to occur when the receiver population is less responsive to the sender population.

Given that the evolution of signalling is supposed to be driven by adaptation, this suggests that the evolution of such an asymmetry (or signalling systems which take advance of asymmetric realiser mechanisms) would itself be advantageous. This is relevant to the ‘compatibility’ signalling template that was considered for RST in chapter 6, as well as earlier in the current chapter. If evolutionary mechanisms for sender and receiver strategies are grounded in the way suggested, i.e. with cultural evolution playing a far greater role on the sender-side, then coordination-signalling interpretations could be promising options for robust religious signalling. Since the result does not hold for conflict of interest games, the effect would only have a limited role in the evolution of religious signalling. That is, unless the basic rationale with respect to the cooperation problem was misguided, and coordination-like strategic scenarios (including perhaps those represented by stag hunt payoff structures as well) were actually the dominant strategic context for religious signalling. This is a possibility, but not one that will be pursued any further here.

However, there are other, more indirect ways that the sender-led asymmetry effect might play a role with respect to conflict of interest signalling. Interestingly, many scholars in the animal communications literature have noted that such a response asymmetry holds between sender and receiver when the interests of the parties partially conflict. For instance, (Owren, Rendall, and Ryan 2010) note that senders can easily adapt their signalling behaviour while receivers for the most part have responses to the stimuli produced by senders that are more difficult to change. This forms part of the contemporary motivation to see signalling as driven by sender manipulation of the receiver.

Yet this is puzzling. In the David Lewis signalling game, it makes sense for the receiver to 'outsource' the evolution of signalling systems to the sender, as they have the same interests. But if there is a partial conflict of interest between sender and receiver, what prevents receivers from increasing the speed at which they adapt to the behaviour of the sender? Indeed, the negative results in 8.2.4 show that this stalemate should be expected: senders can manipulate receivers briefly, but receivers will slowly respond and resist the exploitation, meaning that the ultimate equilibria are unchanged. One well-known explanation here is the stable parasite explanation: the receiver has bigger fish to fry with their conditional strategies, and the rate and degree of manipulation amounts to an annoyance rather than a genuinely disruptive problem. Another is evolvability: the exploited receiver has no fitness-improving pathways for remedial adaptation.

However, two further explanations suggest themselves, based on the asymmetry results. The first is that some apparent cases of manipulation may be transient and out-of-equilibrium, with senders having rapidly converged on a manipulative equilibrium due to a much faster evolutionary rate, and receivers still in the process of adapting to them. The second is that senders and receivers who routinely interact rarely find themselves engaged *exclusively* in either common or conflict of interest signalling games. As is well known by any parent, not all signalling interactions between relatives are free of conflict. Likewise, agents whose interests are typically thought to be partially opposed, such as two potential mates, may frequently engage in common interest signalling games in contexts unrelated to mating. The point here is that a variety of distinct strategic scenarios can hold between sender and receiver. This is a problem of evolutionary 'grain': determine the trait that is actually evolving. There is no principled reason to think all interactions will involve the perfect alignment of interests or

sizable conflict, or that communicative strategies will always be fine-grained enough for scenario-by-scenario optimization.

If this is correct, then when a sizable proportion of interactions between sender and receiver involve no or very low conflict of interest, the generational asymmetry result from the Lewis game may hold to some degree. This will ensure that both sender and receiver profit when the sender population evolves at a faster rate than the receiver population. In other words, receivers in this context do best to limit how responsive they are to senders, to ensure the emergence of informative signalling systems in those cases where their interests do overlap.

For the purposes of RST, the signalling explanation might therefore be given more credence by expanding the scope of signalling beyond purely religious signalling. What other forms of cooperative signalling might also be served by the same faculties and proclivities which underpin religious sender and receiver strategies? Without filling in the mechanistic gaps here we can only speculate, but they might include mate display and other forms of social signalling, language, and other deeply entrenched social practices. In a sense, this move would transform religious signalling (per se) into a maladaptive phenomenon: the by-product of other, more adaptive signalling behaviours.

8.3.3. Exotic equilibria and costly signals as fitness traps

The final set of results suggest a more dramatic departure from the standard use and interpretation of signalling models. Even in parameter spaces where classic separating equilibria are ruled out, a surprisingly high proportion of agents might be engaging in costly signalling behaviour.

In the handicap signalling game, when high types are rare, this is due to hybrid equilibria where signalling strategies are not driven to fixation but rather oscillate in mixed proportions. Honest high types and a subset of dishonest low types jockey expensively for attention from receivers, who waver in how seriously they take those signals (in response to the composition of those who send them). Since there is no minimum signal cost for low types (above zero), such models are interesting possibilities to investigate the initial stages of the evolution of costly signalling. Regarding the interpretive templates from chapter 6, this would correspond to any one of the differential cost templates, except with their uptake being imperfect and subject to dynamic stability (i.e. oscillating over time).

Costly ‘signals’ are also observed even when they do not count as signals in the technical, signalling-system sense of the word – when fitness traps lock senders into costly strategies because of their transient informational value. While such equilibria are not signalling systems, they are equilibria in signalling games which can stabilise a significant amount of ‘signalling’ behaviour, even when signal costs are significant.

This provides an alternative, *maladaptive* explanatory signalling mechanism which could easily be applied in the religious signalling case: early, misguided experiments in religious signals can become entrenched and then conserved, despite doing no good whatsoever and in fact wasting significant resources, simply due to cultural inertia and conservatism. And the key phenomenon to map over to the real world is conservatism with respect to receiver reactions or expectations with respect to signalling. In the R3 equilibrium that was examined, agents playing the receiver role have no material incentive to switch from demanding the signal as a cost of being nice to simply being nice to everyone. It would be better for the overall population if they did so (including themselves, when they switch to the sender role), but if the realisers of sender and receiver strategies (and their update mechanics) are separate, their needs as senders need not impinge on their behaviour as receivers. And if there were some other mechanism like a conformity bias or group fusion in play, there might be considerable resistance to relaxing maladaptive expectations for the costly signal. This schematic outline of an explanation would of course need to be fleshed out more (with the receiver-side interpretation perhaps having testable implications for the cognitive science of religion), but is suggestive of an approach that could leverage signalling theory in an explanatory way without being committed to the adaptive benefits of costly signalling.

This demonstrates both an unexpected opportunity for RST and an interesting conceptual divergence. While signalling theories of religion are often classified as selectionist or adaptive (see section 1.2.2), the hybrid and pooling equilibria investigated here show that explanations based on costly signalling are potentially more versatile than this, especially if transitory information is considered. If reasonable ‘fitness trap’ interpretations can be made with respect to religious signalling, then favourable explanatory work can potentially be carried out where no real cooperative benefits appear to be derived from signalling traditions. This is a class of behaviour (all cost, no benefit) which had been previously left up to maladaptive or cultural-evolutionary mechanistic explanations such as CREDs to explain. With separating, hybrid, and fitness trap equilibria in mind though, religious signalling could be i) adaptive, honest and

stable over time, ii) partially adaptive, partially honest and dynamic over time, or iii) useless, ultimately maladaptive, but driven to fixation by individual selection nevertheless. Each of these might be plausible in different contexts, for different modes of religious expression.

In any case, better exploration of non-standard equilibria in signalling games has the potential to do complementary explanatory work with respect to traditional RST and other explanatory approaches.

8.4. Limitations and further questions

These results are all suggestive; and demonstrate the sort of work that formal modelling can do with regard to exploring the relative merits of signalling explanations. The exact predictions and applicability of a signalling model is not always transparent, and simulations have the potentially validate or falsify hypotheses, map out parameter-spaces for interpretation to real-world target systems, and reveal surprising explanatory options. But the results reported here are of course subject to caveats and limitations.

To begin with, they are simulations based on very simple models. Other than the hedgehog strategy game, they use 2x2x2 signalling models, and some features of the results (such as the exact shape of the curve in figure 8-2) are obvious artefacts of this limited dimensionality. Others might be less obviously so. For example, a general exploration of games with more world-states, signals and/or actions (for example allowing partial-pooling equilibria) would help reveal which features of these results are robust in the face of such de-idealisation. Moving from continuous population to finite population (or agent-based) models would also add a layer of realism for replicating (or not) the general findings.

As the most commonly cited and discussed signalling games, I have also concentrated on the David Lewis and differential intrinsic cost signalling games for these simulations. Similar results were apparent in the differential vulnerability cost signalling game, and the differential benefit game is expected to have results which mirror differential intrinsic cost signalling. However, a more thorough exploration of these alternative differential cost-benefit games (and the whole class of such games) is also warranted. Also lacking was any investigation of index/constraint models. Despite its relative neglect in the literature, index signalling is perfectly capable of being modelled alongside differential cost-benefit games (Simon M. Huttegger, Bruner, and Zollman 2015).

Another limitation is with the method of exogenous parameterisation: costs and probabilities were set and held fixed during each simulation run. The obvious next step would be replicate these results in models where the proportion of high type senders (for example) was treated as a variable rather than a parameter and allowed to vary (for example by differential fitness of types). This could address some of the gaps identified in the literature in chapter 7, with respect to better validating religious signalling as a co-evolutionary theory. The evolution of signal cost too would be valuable to investigate endogenously, perhaps in concert with the endogenous evolution of signal form, to inform (for example) the relative likelihood of differential cost vs costly index signalling.

Finally, I have not demonstrated any robust conclusions about fitness traps here. As already noted, the result has only been demonstrated in one signalling game, with one set of parameters. This means that more work would have to be done to be sure that the fitness-trap result wasn't just an artefact of modelling simplifications. Furthermore, the fitness trap equilibrium in this result is only *weakly* stable, in the sense that the conditional receiver strategy which entraps the senders is no longer being actively stabilised. If it were to drift or decay (for example by being slightly more expensive to maintain than an unconditional strategy, which is not implausible), then the fitness trap would decay as well. The plausibility of fitness traps in this context therefore depends on further interpretive plausibility: whether receiver-strategy inertia would be as enough to stabilise a costly signalling fitness trap, as supposed. This would depend on what else is in it for the receiver (if part of the signal cost includes a windfall gain for receivers then it would be far more robust), conformity and other biases, and the degree to which the waste of signal costs impact on the group's common goods or inter-group competitiveness. It should also depend on cultural drift, mutation, migration, etc. This of course connects back to the syncretic, narrative approach of (Sterelny 2007), where fitness traps play one part in a putative explanation of maladaptive cultural traits. Such factors might be key to whether this sort of result has any bearing on the explanatory virtues of signalling theory in this context, and the reality here (in both world and model) has potential to surprise.

There is, in other words, no shortage of RST-relevant work to do both with respect to formal modelling and to the interpretation of signalling models back onto human societies as target systems.

9. Prospects and outstanding issues

Time then to review where we have ended up, and how we got here. The overall approach to this project has been to explore religious signalling theory mechanistically, by focusing on the building blocks of discrete, potentially explanatory signalling mechanisms using contemporary signalling theory, models and formal methods. The groundwork for this was laid in chapter one, where a landscape of evolutionary approaches to cooperation and religion was mapped out, as well as evidence from the archaeological record that made a prima face case for linking the evolution of cooperation with the evolution of religion. Religion and cooperation are puzzlingly costly in most contexts, but the Holocene transition, I argued, is as good a problem as we are ever likely to find such that a co-evolutionary theory of religion and cooperation (like RST) might be its solution.

In that chapter, I also explained what it is to focus on mechanisms vs big-picture theories, and how these two levels of explanation interact and inform one another. The case study in chapter two, the Big Gods theory, was a demonstration of this, and in assessing the Big Gods theory as a mechanistic complex I introduced considerations (such as parsimony considerations) and distinctions with relevance to the later chapters. I concluded that the Big Gods co-evolutionary approach was promising and intriguing, with an appropriate degree of scope and sophistication. But it is beset (at least for now) with questions regarding the empirical adequacy of both its big-picture predictions and its psychological-mechanistic details.

Chapter three defined a core notion of RST that has at least the potential to avoid such empirical pitfalls, and it exhibits a number of attractive explanatory virtues. In effect, I abstracted away Core RST (as I called it) from the more mixed-mechanistic approaches to RST in the literature (which e.g. also include psychological commitments and other, non-signalling causal mechanisms), and isolated it as a single, functionally-defined mechanistic *schema*. The central concepts of signalling theory itself (signalling systems, manipulation, payoff-sensitivity, etc) were introduced in chapter four, as well as the broad families of signalling models which fit this schema: hard to fake index signals, and fakeable differential cost-benefit signals. I argued that signal costs can be a red-herring, and a potentially deceptive basis for basing investigations into religious signalling. Several examples of signalling and ‘signalling’ were compared and contrasted.

Formal models were introduced in chapter five: the David Lewis signalling game, the differential cost game, and the vulnerability cost game. These were explained within the context of the sender-receiver framework, their game-theoretic rationale, and in comparison to their more general forms and place within the literature (for example in contrast to the Sir Phillip Sydney game). I argued that it is extremely easy to over-interpret the standard handicap signalling game, and that (as an example of this) the stotting example that has been used by so many as a model for the handicap signalling game (despite already conflicting interpretations as an index signal) is in fact better modelled as a case of vulnerability signalling. This discussion hinged on how costs are interpreted, showing that i) nuances of cost-benefit interpretation can make a significant difference with respect to modelling choices, and ii) modelling choices can make an even more significant difference with respect to the evolutionary behaviour and predictions that result.

This theme was continued in chapter six when considering payoff-update systems: costs and benefits must be not just interpreted accurately, but also conceptualised alongside a selection-like update principle, as part of a dynamic system which drives the evolution of sender and receiver strategies. This is what signalling theories of religion are committed to by virtue of appealing to signalling theory, so they must be a coherent story to tell about this (fully specified or not). I then questioned the level of detail at which some of the literature on RST carries out the discussion of signalling and its application to religion. I argued that while it is suitable for broad theorising, there is a gap to be filled between that theorising and generating testable hypotheses. The chapter concluded with some suggested application ‘templates’ for starting to bridge this gap and applying specific signalling models to certain classes of strategic setting: from in-the-moment rituals to long-term interactions between communities and their members. This, and arguments along the way, showed that RST can afford to be much broader and disjunctive in its approach: there are many types of signalling for many types of situations, and great potential (though much work to do) because of this pluripotency diversity. This is the main lesson of the dissertation with respect to RST itself: the approach has both greater potential and greater complexity than we might otherwise think.

Chapter seven turned toward more specific questions that formal modelling might address. How can we model religion-cooperation co-evolution directly, rather than just suggestively via modelling the evolution of signalling within a cooperation-problem framework? What happens to this in different strategic environments (e.g. coordination or stag-hunt environments)? What

can we say about the evolution of signalling form? What can we say about how abstract sender-receiver strategies are realised? What does the fact that they might be realised differently mean for the prospects of the evolution of religious signalling, and therefore RST as an explanation? Various ways of answering this last question were explored in chapter eight, where simulation results showed that costless signalling games are even more likely under the conditions supposed: where sender-side attentiveness and adaptation means that senders update their strategies faster than receivers. However, there is no difference in this regard for costly signalling – at least not directly. Several routes for indirect influence (bleed-over, grain problems, etc) were argued for. But the analysis here also went beyond separating equilibria robustness analysis: hybrid and pooling equilibria both offer tantalising possibilities for explaining the prevalence of costly behaviour among populations where signalling is possible, even when the actual levels of cooperatively useful information are volatile or non-existent in the long run.

While the formal methods used in this dissertation were simple and had important limitations, at the very least they serve as both a demonstration of the potential utility of RST-focused formal modelling and a starting point for more sophisticated iterations of it. This is the more methodological point of the thesis: mixed methods in the science of religion can inform and enrich one another and there is a wealth of possibility for RST with respect to integrating more sophisticated formal modelling projects into the expanding literature.

There is one final lacuna to briefly acknowledge. Nothing about signalling theory necessitates that it must have been *religion* that realised the functional role of an evolved signal. We have been assuming that religion supplies the signal content for commitment signalling models, but signalling models are of course more general than this and could equally apply to secular displays, rituals, paywalls, etc. The apparent ubiquity of religion not fully explained by signalling mechanisms alone, and an additional story beyond that given here will presumably be needed.

But this is to be expected. Big-picture theorising has not been the project here: I have been focusing purely on the core RST mechanism and how best to make that work. Nor have I considered how religious signalling, and characteristically religious beliefs, emotions and values interrelate with signalling and with one another. Religious signalling and religious signalling theory qua stand-alone explanation are not synonymous. Incorporating a plausible

selection of other mechanisms back into an RST-centric big picture theory might well result in something more satisfactory in this regard. Alternatively, the best role for signalling mechanisms might be playing second fiddle to other, more dominantly explanatory mechanisms. Either way, the prospects for RST mechanisms and applications look promising in the science of religion and in the broader project of understanding our social and cultural origins.

Exactly how rosy that future might be depends on how researchers respond to the trade-offs inherent in this approach. Producing useful, testable hypotheses about signalling mechanisms will be a lot of work and require considerable investment in mixed methods approaches and modelling at different levels of generality and realism. So, while RST is perfectly respectable as it is, taking it further will be a matter of delving more rigorously and comprehensively into the messy details. Because the devil is in the details.

BIBLIOGRAPHY

Abbot, Patrick, Jun Abe, John Alcock, Samuel Alizon, Joao A. C. Alpedrinha, Malte Andersson, Jean-Baptiste Andre, et al. 2011. “Inclusive Fitness Theory and Eusociality.” *Nature* 471 (7339): E1–4. https://doi.org/10.1038/nature09831.

Ahmed, Ali M. 2009. “Are Religious People More Prosocial? A Quasi-Experimental Study with ‘Madrasah’ Pupils in a Rural Community in India.” *Journal for the Scientific Study of Religion* 48 (2): 368–74.

Alcorta, Candace. 2017. “Religion, Social Signaling, and Health: A Psychoneuroimmunological Approach.” *Religion, Brain & Behavior* 7 (3): 243–46. https://doi.org/10.1080/2153599X.2016.1156559.

Alcorta, Candace S., and Richard Sosis. 2005. “Ritual, Emotion, and Sacred Symbols : The Evolution of Religion as an Adaptive Complex.” *Human Nature* 16 (4): 323–59. https://doi.org/10.1007/s12110-005-1014-3.

Alexander, Richard D. 1987. *The Biology of Moral Systems*. Foundations of Human Behavior. Hawthorne, N.Y: A. de Gruyter.

Arranz-Otaegui, Amaia, Lara Gonzalez Carretero, Monica N. Ramsey, Dorian Q. Fuller, and Tobias Richter. 2018. “Archaeobotanical Evidence Reveals the Origins of Bread 14,400 Years Ago in Northeastern Jordan.” *Proceedings of the National Academy of Sciences*, July, 201801071. https://doi.org/10.1073/pnas.1801071115.

Atkinson, Quentin D., Andrew J. Latham, and Joseph Watts. 2014. “Are Big Gods a Big Deal in the Emergence of Big Groups?” *Religion, Brain & Behavior*, July, 1–9. https://doi.org/10.1080/2153599X.2014.928351.

Atran, Scott. 2002. *In Gods We Trust: The Evolutionary Landscape of Religion*. Oxford; New York: Oxford University Press.

Atran, Scott, and Joseph Henrich. 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.

Axelrod, Robert. 1980a. “Effective Choice in the Prisoner’s Dilemma.” *Journal of Conflict Resolution* 24 (1): 3–25. https://doi.org/10.1177/002200278002400101.

———. 1980b. “More Effective Choice in the Prisoner’s Dilemma.” *Journal of Conflict Resolution* 24 (3): 379–403. https://doi.org/10.1177/002200278002400301.

Ayasse, Manfred, Florian P. Schiestl, Hannes F. Paulus, Fernando Ibarra, and Wittko Francke. 2003. “Pollinator Attraction in a Sexually Deceptive Orchid by Means of Unconventional Chemicals.” *Proceedings of the Royal Society of London B: Biological Sciences* 270 (1514): 517–22. https://doi.org/10.1098/rspb.2002.2271.

Bandes, Susan A. 2016. “Remorse and Criminal Justice.” *Emotion Review* 8 (1): 14–19. https://doi.org/10.1177/1754073915601222.

Banning, E. B. 2011. “So Fair a House: Göbekli Tepe and the Identification of Temples in the Pre-Pottery Neolithic of the Near East.” *Current Anthropology* 52 (5): 619–60. https://doi.org/10.1086/661207.

Barker, Graeme. 2006. *The Agricultural Revolution in Prehistory: Why Did Foragers Become Farmers?* Oxford ; New York: Oxford University Press.

Barrett, Justin L. 2000. “Exploring the Natural Foundations of Religion.” *Trends in Cognitive Sciences* 4 (1): 29–34. https://doi.org/10.1016/S1364-6613(99)01419-9.

Barrett, Matthew, and Peter Godfrey-Smith. 2002. “Group Selection, Pluralism, and the Evolution of Altruism.” *Philosophy and Phenomenological Research* 65 (3): 685–91. https://doi.org/10.1111/j.1933-1592.2002.tb00233.x.

Bateson, M., D. Nettle, and G. Roberts. 2006. “Cues of Being Watched Enhance Cooperation in a Real-World Setting.” *Biology Letters* 2 (3): 412–14. https://doi.org/10.1098/rsbl.2006.0509.

Batson, C. Daniel, Patricia Schoenrade, and W. Larry Ventis. 1993. *Religion and the Individual: A Social-Psychological Perspective*. Revised edition. New York: Oxford University Press.

Baumard, Nicolas, and Pascal Boyer. 2014. “Empirical Problems with the Notion of ‘Big Gods’ and of Prosociality in Large Societies.” *Religion, Brain & Behavior*, July, 14–18. https://doi.org/10.1080/2153599X.2014.928349.

Beaumont, Peter B., and Robert G. Bednarik. 2013. “Tracing the Emergence of Palaeoart in Sub-Saharan Africa.” *Rock Art Research: The Journal of the Australian Rock Art Research Association (AURA)* 30 (1): 33.

Bednarik, Robert G. 2003. “A Figurine from the African Acheulian.” *Current Anthropology* 44 (3): 405–13. https://doi.org/10.1086/374900.

Bellah, Robert N. 2011. *Religion in Human Evolution: From the Paleolithic to the Axial Age*. Cambridge, Mass: Belknap Press of Harvard University Press.

Bennett, Albert F., and R. B. Huey. 1990. “Studying the Evolution of Physiological Performance.” *Oxford Surveys in Evolutionary Biology* 7: 251–284.

Bergson, Henri. 1932. *Les deux sources de la morale et de la religion*. Paris: Félix Alcan.

Berwick, Robert C., and Noam Chomsky. 2016. *Why Only Us: Language and Evolution*. Cambridge, MA: The MIT Press.

Biernaskie, J. M., A. Grafen, and J. C. Perry. 2014. “The Evolution of Index Signals to Avoid the Cost of Dishonesty.” *Proceedings of the Royal Society B: Biological Sciences* 281 (1790): 20140876–20140876. https://doi.org/10.1098/rspb.2014.0876.

Biernaskie, Jay M., Jennifer C. Perry, and Alan Grafen. 2018. “A General Model of Biological Signals, from Cues to Handicaps.” *Evolution Letters* 2 (3): 201–9. https://doi.org/10.1002/evl3.57.

Birch, Jonathan. 2016. “Hamilton’s Two Conceptions of Social Fitness.” *Philosophy of Science* 83 (5): 848–60. https://doi.org/10.1086/687869.

———. 2017. *The Philosophy of Social Evolution*. Oxford, New York: Oxford University Press.

Birch, Jonathan, and Samir Okasha. 2015. “Kin Selection and Its Critics.” *BioScience* 65 (1): 22–32. https://doi.org/10.1093/biosci/biu196.

Bloch, Maurice. 2008. “Why Religion Is Nothing Special but Is Central.” *Philosophical Transactions of the Royal Society B: Biological Sciences* 363 (1499): 2055–61. https://doi.org/10.1098/rstb.2008.0007.

Boehm, Chris. 1999. *Hierarchy in the Forest*. Cambridge, Mass: Harvard University Press.

Bogucki, Peter I. 1999. *The Origins of Human Society*. The Blackwell History of the World. Malden, Mass. ; Oxford: Blackwell Publishers.

Boster, James S., James Yost, and Catherine Peeke. 2003. “Rage, Revenge, and Religion: Honest Signaling of Aggression and Nonaggression in Waorani Coalitional Violence.” *Ethos* 31 (4): 471–94.

Bourrat, Pierrick. 2015. “Origins and Evolution of Religion from a Darwinian Point of View: Synthesis of Different Theories.” In *Handbook of Evolutionary Thinking in the Sciences*, edited by Thomas Heams, Philippe Huneman, Guillaume Lecointre, and Marc Silberstein, 761–80. Dordrecht: Springer Netherlands. http://link.springer.com/10.1007/978-94-017-9014-7_36.

Bowles, Samuel, Robert Boyd, Sarah Mathew, and Peter J. Richerson. 2012. “The Punishment That Sustains Cooperation Is Often Coordinated and Costly.” *Behavioral and Brain Sciences* 35 (01): 20–21. https://doi.org/10.1017/S0140525X1100118X.

Bowles, Samuel, and Herbert Gintis. 2011. *A Cooperative Species: Human Reciprocity and Its Evolution*. Princeton: Princeton University Press.

Boyd, Richard. 1999. “Homeostasis, Species, and Higher Taxa.” In *Species: New Interdisciplinary Essays*, edited by R. A. Wilson, 141–85. MIT Press.

Boyd, Richard N. 1999. “Kinds, Complexity and Multiple Realization: Comments on Millikan’s ‘Historical Kinds and the Special Sciences.’” *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 95 (1/2): 67–98.

Boyd, Robert, Herbert Gintis, and Samuel Bowles. 2010. “Coordinated Punishment of Defectors Sustains Cooperation and Can Proliferate When Rare.” *Science* 328 (5978): 617–20. https://doi.org/10.1126/science.1183665.

Boyd, Robert, Herbert Gintis, Samuel Bowles, and Peter J. Richerson. 2003. “The Evolution of Altruistic Punishment.” *Proceedings of the National Academy of Sciences* 100 (6): 3531–35. https://doi.org/10.1073/pnas.0630443100.

Boyd, Robert, and Peter J. Richerson. 2009. “Culture and the Evolution of Human Cooperation.” *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 364 (1533): 3281–88. https://doi.org/10.1098/rstb.2009.0134.

Boyer, Pascal. 2001. *Religion Explained: The Evolutionary Origins of Religious Thought*. New York: Basic Books.

Boyer, Pascal, and Nicolas Baumard. 2016. “Projecting WEIRD Features on Ancient Religions.” *Behavioral and Brain Sciences* 39 (January): e6. https://doi.org/10.1017/S0140525X15000369.

Boyer, Pascal, and Pierre Liénard. 2006. “Why Ritualized Behavior? Precaution Systems and Action Parsing in Developmental, Pathological and Cultural Rituals.” *Behavioral and Brain Sciences* 29 (6): 595–613. https://doi.org/10.1017/S0140525X06009332.

Brown, Rachael L. 2014. “What Evolvability Really Is.” *The British Journal for the Philosophy of Science* 65 (3): 549–72. https://doi.org/10.1093/bjps/axt014.

Bruner, Justin, Cailin O’Connor, Hannah Rubin, and Simon M. Huttegger. 2014. “David Lewis in the Lab: Experimental Results on the Emergence of Meaning.” *Synthese*, September, 1–19. https://doi.org/10.1007/s11229-014-0535-x.

Bruner, Justin P. 2015a. “Disclosure and Information Transfer in Signaling Games.” *Philosophy of Science* 82 (4): 649–66. https://doi.org/10.1086/683016.

———. 2015b. “Diversity, Tolerance, and the Social Contract.” *Politics, Philosophy & Economics* 14 (4): 429–48. https://doi.org/10.1177/1470594X14560763.

———. 2018. “Coordination, Conflict, and Externalization.” *Behavioral and Brain Sciences* 41. https://doi.org/10.1017/S0140525X18000043.

Bruner, Justin P., Carl Brusse, and David Kalkman. 2017. “Cost, Expenditure and Vulnerability.” *Biology & Philosophy* 32 (3): 357–75. https://doi.org/10.1007/s10539-017-9563-5.

Brusse, Carl. 2016. “Planets, Pluralism, and Conceptual Lineage.” *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 53 (February): 93–106. https://doi.org/10.1016/j.shpsb.2015.11.002.

———. 2017. “Making Do without Selection—Review Essay of ‘Cultural Evolution: Conceptual Challenges’ by Tim Lewens.” *Biology & Philosophy* 32 (2): 307–19. https://doi.org/10.1007/s10539-016-9560-0.

Brusse, Carl, and Justin Bruner. 2017. “Responsiveness and Robustness in the David Lewis Signaling Game.” *Philosophy of Science* 84 (5): 1068–79. https://doi.org/10.1086/694156.

Brusse, Carl Joseph, and Kim Sterelny. 2018. “Moral Externalisation Fails to Scale.” *Behavioral and Brain Sciences* 41: e100. https://doi.org/10.1017/S0140525X18000055.

Bulbulia, Joseph. 2004a. “Religious Costs as Adaptations That Signal Altruistic Intention.” *Evolution and Cognition* 10 (1): 19–38.

———. 2004b. “The Cognitive and Evolutionary Psychology of Religion.” *Biology and Philosophy* 19 (5): 655–86. https://doi.org/10.1007/s10539-005-5568-6.

———. 2010. “Charismatic Signalling.” *Journal for the Study of Religion, Nature and Culture* 3 (4). https://doi.org/10.1558/jsrnc.v3i4.518.

———. 2011. “Spreading Order: Religion, Cooperative Niche Construction, and Risky Coordination Problems.” *Biology & Philosophy* 27 (1): 1–27. https://doi.org/10.1007/s10539-011-9295-x.

———. 2013. “Why ‘Costly Signalling’ Models of Religion Require Cognitive Psychology.” In *Origins of Religion, Cognition and Culture*, edited by Armin W. Geertz. Religion, Cognition and Culture. Acumen Publishing.

Bulbulia, Joseph, Gloria Fraser, Joseph Watts, John H. Shaver, and Russell Gray. 2017. “Can Honest Signaling Theory Clarify Religion’s Role in the Evolution of Social Inequality?” *Religion, Brain & Behavior* 7 (4): 285–88. https://doi.org/10.1080/2153599X.2016.1249914.

Bulbulia, Joseph, Armin W. Geertz, Quentin D. Atkinson, Emma Cohen, Nicholas Evans, Pieter François, Herbert Gintis, et al. 2013. “The Cultural Evolution of Religion.” In *Cultural Evolution: Society, Technology, Language, and Religion*, edited by Peter J Richerson and Morten H Christiansen. MIT Press. http://site.ebrary.com/id/10791805.

Bulbulia, Joseph, and Richard Sosis. 2011. “Signalling Theory and the Evolution of Religious Cooperation.” *Religion* 41 (3): 363–88. https://doi.org/10.1080/0048721X.2011.604508.

Buss, David M. 2016. *The Evolution of Desire: Strategies of Human Mating*. 4th ed. Basic Books.

Calcott, Brett. 2008. “The Other Cooperation Problem: Generating Benefit.” *Biology & Philosophy* 23 (2): 179–203. https://doi.org/10.1007/s10539-007-9095-5.

Caro, T. M. 1986a. “The Functions of Stotting: A Review of the Hypotheses.” *Animal Behaviour* 34 (3): 649–62. https://doi.org/10.1016/S0003-3472(86)80051-3.

———. 1986b. “The Functions of Stotting in Thomson’s Gazelles: Some Tests of the Predictions.” *Animal Behaviour* 34 (3): 663–84. https://doi.org/10.1016/S0003-3472(86)80052-5.

Centorrino, Samuele, Elodie Djemai, Astrid Hopfensitz, Manfred Milinski, and Paul Seabright. 2015. “A Model of Smiling as a Costly Signal of Cooperation Opportunities.” *Adaptive Human Behavior and Physiology* 1 (3): 325–40. https://doi.org/10.1007/s40750-015-0026-4.

Clarke, Ellen. 2014. “Origins of Evolutionary Transitions.” *Journal of Biosciences* 39 (2): 303–17. https://doi.org/10.1007/s12038-013-9375-y.

Cressman, Ross, and Yi Tao. 2014. “The Replicator Equation and Other Game Dynamics.” *Proceedings of the National Academy of Sciences* 111 (Supplement 3): 10810–17. https://doi.org/10.1073/pnas.1400823111.

Cronk, Lee. 1994. “Evolutionary Theories of Morality and the Manipulative Use of Signals.” *Zygon*® 29 (1): 81–101. https://doi.org/10.1111/j.1467-9744.1994.tb00651.x.

Darley, John M., and C. Daniel Batson. 1973. “‘From Jerusalem to Jericho’: A Study of Situational and Dispositional Variables in Helping Behavior.” *Journal of Personality and Social Psychology* 27 (1): 100–108.

Davis, J. W. F., and P. O’Donald. 1976. “Sexual Selection for a Handicap: A Critical Analysis of Zahavi’s Model.” *Journal of Theoretical Biology* 57 (2): 345–54. https://doi.org/10.1016/0022-5193(76)90006-0.

Davis, Taylor. 2015. “Group Selection in the Evolution of Religion: Genetic Evolution or Cultural Evolution?” *Journal of Cognition and Culture* 15 (3–4): 235–53. https://doi.org/10.1163/15685373-12342149.

Davis, Taylor, and Daniel Kelly. 2018. “Norms, Not Moral Norms: The Boundaries of Morality Do Not Matter.” *Behavioral and Brain Sciences* 41. https://doi.org/10.1017/S0140525X18000067.

Dawes, Gregory W., and James Maclaurin, eds. 2013. *A New Science of Religion*. Routledge Studies in Religion 23. New York: Routledge.

Dawkins, R., and J. R. Krebs. 1979. “Arms Races between and within Species.” *Proceedings of the Royal Society of London B: Biological Sciences* 205 (1161): 489–511. https://doi.org/10.1098/rspb.1979.0081.

Dawkins, Richard. 1976. *The Selfish Gene*. New York: Oxford University Press.

Dawkins, Richard, and John R Krebs. 1978. “Animal Signals: Information or Manipulation?” In *Behavioural Ecology: An Evolutionary Approach*, edited by John R Krebs and Nicholas B Davies, 282–309. Oxford: Blackwell Scientific.

Dear, Keith, Kevin Dutton, and Elaine Fox. 2019. “Do ‘Watching Eyes’ Influence Antisocial Behavior? A Systematic Review & Meta-Analysis.” *Evolution and Human Behavior* 40 (3): 269–80. https://doi.org/10.1016/j.evolhumbehav.2019.01.006.

Deem, Michael J., and Grant Ramsey. 2016. “Guilt by Association?” *Philosophical Psychology* 29 (4): 570–85. https://doi.org/10.1080/09515089.2015.1126706.

Diamond, Jared. 1991. *The Rise and Fall of the Third Chimpanzee*. London: Vintage.

Diamond, Jared M. 1998. *Guns, Germs and Steel: A Short History of Everybody for the Last 13,000 Years*. Random House.

Dietrich, O., and K. Schmidt. 2010. “A Radiocarbon Date from the Wall Plaster of Enclosure D of Göbekli Tepe.” *Neo-Lithics* 10 (2): 82–3.

Dietrich, Oliver, Manfred Heun, Jens Notroff, Klaus Schmidt, and Martin Zarnkow. 2012. “The Role of Cult and Feasting in the Emergence of Neolithic Communities. New Evidence from Göbekli Tepe, South-Eastern Turkey.” *Antiquity* 86 (333): 674–695.

Dietrich, Oliver, Çiğdem Köksal-Schmidt, Jens Notroff, and Klaus Schmidt. 2013. “Establishing a Radiocarbon Sequence for Göbekli Tepe. State of Research and New Data.” *Neo-Lithics* 13 (1): 36–47.

Dietrich, Oliver, Cigdem Koksal-Schmidt, Jens Notroff, Klaus Schmidt, and Cihat Kurkcuoglu. 2012. “Göbekli Tepe: A Stone Age Ritual Center in Southeastern Turkey.” *Actual Archeology*, 2012.

Diggle, S. P, A. Gardner, S. A West, and A. S Griffin. 2007. “Evolutionary Theory of Bacterial Quorum Sensing: When Is a Signal Not a Signal?” *Philosophical Transactions of the Royal Society B: Biological Sciences* 362 (1483): 1241–49. https://doi.org/10.1098/rstb.2007.2049.

Dirks, Paul HGM, Lee R Berger, Eric M Roberts, Jan D Kramers, John Hawks, Patrick S Randolph-Quinney, Marina Elliott, et al. 2015. “Geological and Taphonomic Context for the New Hominin Species Homo Naledi from the Dinaledi Chamber, South Africa.” Edited by Nicholas J Conard and Johannes Krause. *ELife* 4 (September): e09561. https://doi.org/10.7554/eLife.09561.

Durand, François. 2017. “Naledi: An Example of How Natural Phenomena Can Inspire Metaphysical Assumptions.” *HTS Theological Studies* 73 (3): 1–9. https://doi.org/10.4102/hts.v73i3.4507.

Durkheim, E. C. 1912. *The Elementary Forms of the Religious Life, Trans. Karen Fields*. New York: Free Press.

Fischer, Ronald, and Dimitris Xygalatas. 2014. “Extreme Rituals as Social Technologies.” *Journal of Cognition and Culture* 14 (5): 345–55. https://doi.org/10.1163/15685373-12342130.

Fisher, R. A. 1930. *The Genetical Theory of Natural Selection: A Complete Variorum Edition*. OUP Oxford.

FitzGibbon, C. D., and J. H. Fanshawe. 1988. “Stotting in Thomson’s Gazelles: An Honest Signal of Condition.” *Behavioral Ecology and Sociobiology* 23 (2): 69–74. https://doi.org/10.1007/BF00299889.

Foley, Robert. 1988. “Hominids, Humans and Hunter-Gatherers: An Evolutionary Perspective.” In *Hunters and Gatherers 1: History, Evolution and Social Change*, edited by Tim Ingold, David Riches, and James Woodburn, 1:207–221. Oxford: Berg.

Forber, Patrick. 2010. “Confirmation and Explaining How Possible.” *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 41 (1): 32–40. https://doi.org/10.1016/j.shpsc.2009.12.006.

Forber, Patrick, and Rory Smead. 2014a. “The Evolution of Fairness through Spite.” *Proceedings of the Royal Society B: Biological Sciences* 281 (1780): 20132439. https://doi.org/10.1098/rspb.2013.2439.

———. 2014b. “An Evolutionary Paradox for Prosocial Behavior.” *The Journal of Philosophy* 111 (3): 151–66. https://doi.org/10.5840/jphil201411139.

———. 2015. “Evolution and the Classification of Social Behavior.” *Biology & Philosophy* 30 (3): 405–21. https://doi.org/10.1007/s10539-015-9486-y.

———. 2016. “The Evolution of Spite, Recognition, and Morality.” *Philosophy of Science* 83 (5): 884–96. https://doi.org/10.1086/687872.

Frank, Robert H. 1988. *Passions Within Reason: The Strategic Role of Emotions*. New York: W W Norton & Co Inc.

Fraser, Ben. 2011. “Costly Signalling Theories: Beyond the Handicap Principle.” *Biology & Philosophy* 27 (2): 263–78. https://doi.org/10.1007/s10539-011-9297-8.

Freud, Sigmund. 1927. *Die Zukunft Einer Illusion*. Leipzig, Wien und Zürich: Internationaler Psychoanalytischer Verlag.

Gall, Robert S. 1999. “Kami and Daimōn: A Cross-Cultural Reflection on What Is Divine.” *Philosophy East and West* 49 (1): 63–74. https://doi.org/10.2307/1400117.

Gardner, A., and A. Grafen. 2009. “Capturing the Superorganism: A Formal Theory of Group Adaptation.” *Journal of Evolutionary Biology* 22 (4): 659–71. https://doi.org/10.1111/j.1420-9101.2008.01681.x.

Gibbons, Robert. 1992. *A Primer in Game Theory*. Harlow: Pearson Higher Education.

Gintis, H., E.A. Smith, and S. Bowles. 2001. “Costly Signaling and Cooperation.” *Journal of Theoretical Biology* 213 (1): 103–19.

Godfrey-Smith, Peter. 1994. “A Modern History Theory of Functions.” *Noûs* 28 (3): 344–62. https://doi.org/10.2307/2216063.

———. 2009. *Darwinian Populations and Natural Selection*. Oxford University Press.

———. 2011. “Signals: Evolution, Learning, and Information, by Brian Skyrms.” *Mind* 120 (480): 1288–97. https://doi.org/10.1093/mind/fzs002.

Godfrey-Smith, Peter, and Manolo Martínez. 2013. “Communication and Common Interest.” *PLoS Comput Biol* 9 (11): e1003282. https://doi.org/10.1371/journal.pcbi.1003282.

Gomes, Cristina M., and Michael E. McCullough. 2015. “The Effects of Implicit Religious Primes on Dictator Game Allocations: A Preregistered Replication Experiment.” *Journal of Experimental Psychology: General*, No Pagination Specified. https://doi.org/10.1037/xge0000027.

Gould, S. J., and R. Lewontin. 1978. “The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme.” *Proceedings of the Royal Society, London (Series B)* 205: 581–98.

Grafen, Alan. 1990a. “Biological Signals as Handicaps.” *Journal of Theoretical Biology* 144 (4): 517–46. https://doi.org/10.1016/S0022-5193(05)80088-8.

———. 1990b. “Sexual Selection Unhandicapped by the Fisher Process.” *Journal of Theoretical Biology* 144 (4): 473–516. https://doi.org/10.1016/S0022-5193(05)80087-6.

Griffiths, Paul E. 1993. “Functional Analysis and Proper Functions.” *The British Journal for the Philosophy of Science* 44 (3): 409–22. https://doi.org/10.1093/bjps/44.3.409.

Grose, Jonathan. 2011. “Modelling and the Fall and Rise of the Handicap Principle.” *Biology & Philosophy* 26 (5): 677–96. https://doi.org/10.1007/s10539-011-9275-1.

Groucutt, Huw S., Michael D. Petraglia, Geoff Bailey, Eleanor M. L. Scerri, Ash Parton, Laine Clark-Balzan, Richard P. Jennings, et al. 2015. “Rethinking the Dispersal of Homo Sapiens out of Africa.” *Evolutionary Anthropology: Issues, News, and Reviews* 24 (4): 149–64. https://doi.org/10.1002/evan.21455.

Guthrie, Stewart Elliott. 1995. *Faces in the Clouds: A New Theory of Religion*. New York: Oxford University Press.

Haidt, Jonathan. 2013. *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. Reprint edition. New York: Vintage.

Hales, David. 2005. “Change Your Tags Fast! – A Necessary Condition for Cooperation?” In *Multi-Agent and Multi-Agent-Based Simulation*, edited by Paul Davidsson, Brian Logan, and Keiki Takadama, 89–98. Lecture Notes in Computer Science. Springer Berlin Heidelberg.

Haley, Kevin J., and Daniel M. T. Fessler. 2005. “Nobody’s Watching?: Subtle Cues Affect Generosity in an Anonymous Economic Game.” *Evolution and Human Behavior* 26 (3): 245–56. https://doi.org/10.1016/j.evolhumbehav.2005.01.002.

Hall, Deborah L., and J. P. Gonzales. 2016. “Religious Group Identity and Costly Signaling.” *Religion, Brain & Behavior* 0 (0): 1–3. https://doi.org/10.1080/2153599X.2016.1156564.

Hamilton, W. D. 1964a. “The Genetical Evolution of Social Behaviour. I.” *Journal of Theoretical Biology* 7 (1): 1–16. https://doi.org/10.1016/0022-5193(64)90038-4.

———. 1964b. “The Genetical Evolution of Social Behaviour. II.” *Journal of Theoretical Biology* 7 (1): 17–52. https://doi.org/10.1016/0022-5193(64)90039-6.

Hamilton, William D., and Marlene Zuk. 1982. “Heritable True Fitness and Bright Birds: A Role for Parasites?” *Science* 218 (4570): 384–387.

Hardin, Garrett. 1968. “The Tragedy of the Commons.” *Science* 162 (3859): 1243–48. https://doi.org/10.1126/science.162.3859.1243.

Harrison, Victoria S. 2006. “The Pragmatics of Defining Religion in a Multi-Cultural World.” *International Journal for Philosophy of Religion* 59 (3): 133–52. https://doi.org/10.1007/s11153-006-6961-z.

Hebets, Eileen A., Andrew B. Barron, Christopher N. Balakrishnan, Mark E. Hauber, Paul H. Mason, and Kim L. Hoke. 2016. “A Systems Approach to Animal Communication.” *Proc. R. Soc. B* 283 (1826): 20152889. https://doi.org/10.1098/rspb.2015.2889.

Hebets, Eileen A., and Daniel R. Papaj. 2005. “Complex Signal Function: Developing a Framework of Testable Hypotheses.” *Behavioral Ecology and Sociobiology* 57 (3): 197–214. https://doi.org/10.1007/s00265-004-0865-7.

Heisler, I. Lorraine, and John Damuth. 1987. “A Method for Analyzing Selection in Hierarchically Structured Populations.” *The American Naturalist* 130 (4): 582–602. https://doi.org/10.1086/284732.

Henley, Tracy B. 2018. “Introducing Göbekli Tepe to Psychology.” *Review of General Psychology* 22 (4): 477–84. https://doi.org/10.1037/gpr0000151.

Henrich, Joseph. 2004. “Cultural Group Selection, Coevolutionary Processes and Large-Scale Cooperation.” *Journal of Economic Behavior & Organization*, Evolution and Altruism, 53 (1): 3–35. https://doi.org/10.1016/S0167-2681(03)00094-5.

———. 2009. “The Evolution of Costly Displays, Cooperation and Religion.” *Evolution and Human Behavior* 30 (4): 244–60. https://doi.org/10.1016/j.evolhumbehav.2009.03.005.

Henrich, Joseph, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, Richard McElreath, et al. 2005. “‘Economic Man’ in Cross-Cultural Perspective: Behavioral Experiments in 15 Small-Scale Societies.” *Behavioral and Brain Sciences* 28 (06): 795–815.

Henrich, Joseph, Steven J. Heine, and Ara Norenzayan. 2010. “The Weirdest People in the World?” *Behavioral and Brain Sciences* 33 (2–3): 61–83. https://doi.org/10.1017/S0140525X0999152X.

Henshilwood, Christopher S., Francesco d’Errico, and Ian Watts. 2009. “Engraved Ochres from the Middle Stone Age Levels at Blombos Cave, South Africa.” *Journal of Human Evolution* 57 (1): 27–47. https://doi.org/10.1016/j.jhevol.2009.01.005.

Hershkovitz, Israel, Gerhard W. Weber, Rolf Quam, Mathieu Duval, Rainer Grün, Leslie Kinsley, Avner Ayalon, et al. 2018. “The Earliest Modern Humans Outside Africa.” *Science* 359 (6374): 456–59. https://doi.org/10.1126/science.aap8369.

Heyes, Celia. 2018. *Cognitive Gadgets: The Cultural Evolution of Thinking*. Cambridge: Harvard University Press.

Higham, James P. 2014. “How Does Honest Costly Signaling Work?” *Behavioral Ecology* 25 (1): 8–11. https://doi.org/10.1093/beheco/art097.

Hill, Kim R., Robert S. Walker, Miran Božičević, James Eder, Thomas Headland, Barry Hewlett, A. Magdalena Hurtado, Frank Marlowe, Polly Wiessner, and Brian Wood. 2011. “Co-Residence Patterns in Hunter-Gatherer Societies Show Unique Human Social Structure.” *Science* 331 (6022): 1286–89. https://doi.org/10.1126/science.1199071.

Hobaiter, Catherine, and Richard W. Byrne. 2014. “The Meanings of Chimpanzee Gestures.” *Current Biology* 24 (14): 1596–1600. https://doi.org/10.1016/j.cub.2014.05.066.

Hodder, Ian. 2014. “Çatalhöyük: The Leopard Changes Its Spots. A Summary of Recent Work.” *Anatolian Studies* 64: 1–22.

Hofbauer, Josef, and Simon M. Huttegger. 2008. “Feasibility of Communication in Binary Signaling Games.” *Journal of Theoretical Biology* 254 (4): 843–49. https://doi.org/10.1016/j.jtbi.2008.07.010.

Hoffman, Bruce, and Gordon H. McCormick. 2004. “Terrorism, Signaling, and Suicide Attack.” *Studies in Conflict & Terrorism* 27 (4): 243–81. https://doi.org/10.1080/10576100490466498.

Hublin, Jean-Jacques, Abdelouahed Ben-Ncer, Shara E. Bailey, Sarah E. Freidline, Simon Neubauer, Matthew M. Skinner, Inga Bergmann, et al. 2017. “New Fossils from Jebel Irhoud, Morocco and the Pan-African Origin of *Homo Sapiens*.” *Nature* 546 (7657): 289–92. https://doi.org/10.1038/nature22336.

Hurd, Peter L., and Magnus Enquist. 2005. “A Strategic Taxonomy of Biological Communication.” *Animal Behaviour* 70 (5): 1155–70. https://doi.org/10.1016/j.anbehav.2005.02.014.

Huttegger, S. M., and K. J. S. Zollman. 2013. “Methodology in Biological Game Theory.” *The British Journal for the Philosophy of Science* 64 (3): 637–58. https://doi.org/10.1093/bjps/axs035.

Huttegger, Simon M. 2007. “Evolution and the Explanation of Meaning.” *Philosophy of Science* 74 (1): 1–27.

Huttegger, Simon M., Justin P. Bruner, and Kevin J. S. Zollman. 2015. “The Handicap Principle Is an Artifact.” *Philosophy of Science* 82 (5): 997–1009. https://doi.org/10.1086/683435.

Iannaccone, Laurence R. 1992. “Sacrifice and Stigma: Reducing Free-Riding in Cults, Communes, and Other Collectives.” *Journal of Political Economy* 100 (2): 271–91.

———. 1994. “Why Strict Churches Are Strong.” *American Journal of Sociology* 99 (5): 1180–1211. https://doi.org/10.1086/230409.

Irons, William. 1996. “Morality, Religion, and Human Nature.” In *Religion and Science: History, Method, and Dialogue*, edited by W. Mark Richardson and Wesley J. Wildman, 375–399. New York: Routledge.

———. 2001. “Religion as a Hard-to-Fake Sign of Commitment.” In *Evolution and the Capacity for Commitment*, edited by Randolph Nesse, 292–309. Russell Sage Foundation.

James, William. 1902. *The Varieties of Religious Experience: A Study in Human Nature : Being the Gifford Lectutes on Natural Religion Delivered at Edinburgh in 1901-1902*. London; New York: Longmans, Green and Co.

Jansen, Vincent A. A., and Minus van Baalen. 2006. “Altruism through Beard Chromodynamics.” *Nature* 440 (7084): 663–66. https://doi.org/10.1038/nature04387.

Jaubert, Jacques, Sophie Verheyden, Dominique Genty, Michel Soulier, Hai Cheng, Dominique Blamart, Christian Burlet, et al. 2016. “Early Neanderthal Constructions Deep in Bruniquel Cave in Southwestern France.” *Nature* advance online publication (May). https://doi.org/10.1038/nature18291.

Jegindø, Else-Marie Elmholdt, Lene Vase, Jens Jegindø, and Armin W. Geertz. 2013. “Pain and Sacrifice: Experience and Modulation of Pain in a Religious Piercing Ritual.” *The International Journal for the Psychology of Religion* 23 (3): 171–87. https://doi.org/10.1080/10508619.2012.759065.

Jensen, Keith. 2010. “Punishment and Spite, the Dark Side of Cooperation.” *Philosophical Transactions of the Royal Society B: Biological Sciences* 365 (1553): 2635–50. https://doi.org/10.1098/rstb.2010.0146.

Jensen, Niels Holm, Michael Bang Petersen, Henrik Høgh-Olesen, and Michael Ejstrup. 2015. “Testing Theories about Ethnic Markers.” *Human Nature* 26 (2): 210–34. https://doi.org/10.1007/s12110-015-9229-4.

Johnson, Dominic. 2015. *God Is Watching You: How the Fear of God Makes Us Human*. 1 edition. New York: Oxford University Press.

Johnson, Dominic, and Jesse Bering. 2006. “Hand of God, Mind of Man: Punishment and Cognition in the Evolution of Cooperation.” *Evolutionary Psychology* 4 (1): 147470490600400130. https://doi.org/10.1177/147470490600400119.

Johnson, Dominic, and Oliver Krüger. 2004. “The Good of Wrath: Supernatural Punishment and the Evolution of Cooperation.” *Political Theology* 5 (2): 159–76. https://doi.org/10.1558/poth.2004.5.2.159.

Johnstone, Rufus A. 1997. “The Evolution of Animal Signals.” In *Behavioural Ecology: An Evolutionary Approach*, edited by John R Krebs and Nicholas B Davies, 155–78. Oxford: Blackwell.

Johnstone, Rufus A., and Alan Grafen. 1992. “The Continuous Sir Philip Sidney Game: A Simple Model of Biological Signalling.” *Journal of Theoretical Biology* 156 (2): 215–34. https://doi.org/10.1016/S0022-5193(05)80674-5.

Joordens, Josephine C. A., Francesco d’Errico, Frank P. Wesselingh, Stephen Munro, John de Vos, Jakob Wallinga, Christina Ankjærgaard, et al. 2015. “*Homo Erectus* at Trinil on Java Used Shells for Tool Production and Engraving.” *Nature* 518 (7538): 228–31. https://doi.org/10.1038/nature13962.

Jordan, Jillian J., Moshe Hoffman, Paul Bloom, and David G. Rand. 2016. “Third-Party Punishment as a Costly Signal of Trustworthiness.” *Nature* 530 (7591): 473–76. https://doi.org/10.1038/nature16981.

Jordan, Jillian J., and David G. Rand. 2017. “Third-Party Punishment as a Costly Signal of High Continuation Probabilities in Repeated Games.” *Journal of Theoretical Biology* 421 (May): 189–202. https://doi.org/10.1016/j.jtbi.2017.04.004.

Jovanovic, Boyan. 1982. “Truthful Disclosure of Information.” *The Bell Journal of Economics* 13 (1): 36. https://doi.org/10.2307/3003428.

Kane, Patrick, and Kevin J. S. Zollman. 2015. “An Evolutionary Comparison of the Handicap Principle and Hybrid Equilibrium Theories of Signaling.” *PLOS ONE* 10 (9): e0137271. https://doi.org/10.1371/journal.pone.0137271.

Kelly, Robert L. 2013. *The Lifeways of Hunter-Gatherers: The Foraging Spectrum*. 2nd ed. Cambridge: Cambridge University Press. https://virtual.anu.edu.au/login/?url=http://www.anu.eblib.com.au/patron/FullRecord.aspx?p=1139695.

Kerr, Benjamin, and Peter Godfrey-Smith. 2002. “Individualist and Multi-Level Perspectives on Selection in Structured Populations.” *Biology and Philosophy* 17 (4): 477–517. https://doi.org/10.1023/A:1020504900646.

Kirschner, Marc, and John Gerhart. 1998. “Evolvability.” *Proceedings of the National Academy of Sciences* 95 (15): 8420–27. https://doi.org/10.1073/pnas.95.15.8420.

Kotiaho, Janne S. 2001. “Costs of Sexual Traits: A Mismatch between Theoretical Considerations and Empirical Evidence.” *Biological Reviews* 76 (3): 365–376.

Krebs, J. R., and R. Dawkins. 1984. “Animal Signals: Mind-Reading and Manipulation.” In *Behavioural Ecology: An Evolutionary Approach*, edited by J. R. Krebs and N. B. Davies, 2nd ed, 380–402. Oxford: Blackwell Scientific.

Lachmann, Michael, Szabolcs Számadó, and Carl T. Bergstrom. 2001. “Cost and Conflict in Animal Signals and Human Language.” *Proceedings of the National Academy of Sciences of the United States of America* 98 (23): 13189–94. https://doi.org/10.1073/pnas.231216498.

Laidre, Mark E., and Rufus A. Johnstone. 2013. “Animal Signals.” *Current Biology* 23 (18): R829–33. https://doi.org/10.1016/j.cub.2013.07.070.

Laland, Kevin N. 2004. “Social Learning Strategies.” *Animal Learning & Behavior* 32 (1): 4–14. https://doi.org/10.3758/BF03196002.

Lamb, Henry F., C. Richard Bates, Charlotte L. Bryant, Sarah J. Davies, Dei G. Huws, Michael H. Marshall, Helen M. Roberts, and Harry Toland. 2018. “150,000-Year Palaeoclimate Record from Northern Ethiopia Supports Early, Multiple Dispersals of Modern Humans from Africa.” *Scientific Reports* 8 (1): 1077. https://doi.org/10.1038/s41598-018-19601-w.

Lapan, Harvey E., and Todd Sandler. 1993. “Terrorism and Signalling.” *European Journal of Political Economy* 9 (3): 383–97. https://doi.org/10.1016/0176-2680(93)90006-G.

Larsen, Clark Spencer. 2006. “The Agricultural Revolution as Environmental Catastrophe: Implications for Health and Lifestyle in the Holocene.” *Quaternary International*, Impact of rapid environmental changes on humans and ecosystems, 150 (1): 12–20. https://doi.org/10.1016/j.quaint.2006.01.004.

Layton, Robert, and Sean O’Hara. 2010. “Human Social Evolution: A Comparison of Hunter-Gatherer and Chimpanzee Social Organization.” In *Social Brain, Distributed Mind*, edited by Robin Dunbar, Clive Gamble, and John Gowlett, 83–113. Proceedings of the British Academy 158. British Academy. http://oxfordindex.oup.com/view/10.5871/bacad/9780197264522.003.0005.

Lee, Richard B, and Irven DeVore, eds. 1968. *Man the Hunter*. New York: Aldine de Gruyter.

Levins, Richard. 1966. “The Strategy of Model Building in Population Biology.” *American Scientist* 54 (4): 421–31.

———. 1968. *Evolution in Changing Environments: Some Theoretical Explorations.* Princeton, NJ: Princeton University Press.

Lewens, Tim. 2015. *Cultural Evolution: Conceptual Challenges*. Oxford University Press.

Lewis, David K. 1969. *Convention: A Philosophical Study,*. Cambridge, MA: Harvard University Press.

Li, Zhan-Yang, Xiu-Jie Wu, Li-Ping Zhou, Wu Liu, Xing Gao, Xiao-Mei Nian, and Erik Trinkaus. 2017. “Late Pleistocene Archaic Human Crania from Xuchang, China.” *Science* 355 (6328): 969–72. https://doi.org/10.1126/science.aal2482.

Liu, Wu, María Martinón-Torres, Yan-jun Cai, Song Xing, Hao-wen Tong, Shu-wen Pei, Mark Jan Sier, et al. 2015. “The Earliest Unequivocally Modern Humans in Southern China.” *Nature* 526 (7575): 696–99. https://doi.org/10.1038/nature15696.

Lyu, Jiankun, Sheng Wang, Trent E. Balius, Isha Singh, Anat Levit, Yurii S. Moroz, Matthew J. O'Meara, et al. 2019. "Ultra-Large Library Docking for Discovering New Chemotypes." *Nature* 566 (7743): 224. https://doi.org/10.1038/s41586-019-0917-9.

Malhotra, Deepak. 2010. "(When) Are Religious People Nicer? Religious Salience and the 'Sunday Effect' on Prosocial Behavior." *Judgment and Decision Making* 5 (2): 143.

Marcus, Joyce. 2008. "The Archaeological Evidence for Social Evolution." *Annual Review of Anthropology* 37 (1): 251–66. https://doi.org/10.1146/annurev.anthro.37.081407.085246.

Martínez, Manolo, and Peter Godfrey-Smith. 2016. "Common Interest and Signaling Games: A Dynamic Analysis." *Philosophy of Science* 83 (3): 371–92. https://doi.org/10.1086/685743.

Matsugasaki, Keisuke, Wakana Tsukamoto, and Yohsuke Ohtsubo. 2015. "Two Failed Replications of the Watching Eyes Effect." *Letters on Evolutionary Behavioral Science* 6 (2): 17–20. https://doi.org/10.5178/lebs.2015.36.

Maynard Smith, John. 1976. "Sexual Selection and the Handicap Principle." *Journal of Theoretical Biology* 57 (1): 239–42. https://doi.org/10.1016/S0022-5193(76)80016-1.

———. 1982a. *Evolution and the Theory of Games*. Cambridge ; New York: Cambridge University Press.

———. 1982b. "Storming the Fortress." *The New York Review of Books*, May 13, 1982. https://www.nybooks.com/articles/1982/05/13/storming-the-fortress/.

———. 1991. "Honest Signalling: The Philip Sidney Game." *Animal Behaviour* 42 (6): 1034–35. https://doi.org/10.1016/S0003-3472(05)80161-7.

Maynard-Smith, John, and David Harper. 2003. *Animal Signals*. Oxford Series in Ecology and Evolution. New York: Oxford University Press.

McElreath, Richard, Robert Boyd, and Peter J. Richerson. 2003. "Shared Norms and the Evolution of Ethnic Markers." *Current Anthropology* 44 (1): 122–30. https://doi.org/10.1086/ca.2003.44.issue-1.

McNamara, Patrick, ed. 2006. *Where God and Science Meet: How Brain and Evolutionary Studies Alter Our Understanding of Religion*. Psychology, Religion, and Spirituality. Westport, Conn: Praeger Publishers.

Millet, Kobe, and Siegfried Dewitte. 2007. "Altruistic Behavior as a Costly Signal of General Intelligence." *Journal of Research in Personality* 41 (2): 316–26. https://doi.org/10.1016/j.jrp.2006.04.002.

Mitchell, Sandra D. 2000. "Dimensions of Scientific Law." *Philosophy of Science* 67 (2): 242–65. https://doi.org/10.1086/392774.

Mithen, S. J. 2004. "From Ohalo to Çatalhöyük: The Development of Religiosity during the Early Prehistory of Western Asia, 20,000-7000 BC." In *Theorizing Religions Past: Archaeology, History, and Cognition*, edited by H. Whitehouse and L. H. Martin, 17–43. Walnut Creek CA: AltaMira Press. http://centaur.reading.ac.uk/3840/.

Mithen, Steven J. 1999. *The Prehistory of the Mind: The Cognitive Origins of Art, Religion and Science*. Thames and Hudson.

Moore, A. M. T, Gordon C Hillman, and A. J Legge. 2000. *Village on the Euphrates: From Foraging to Farming at Abu Hureyra*. London; New York: Oxford University Press.

Morinis, Alan. 1985. "The Ritual Experience: Pain and the Transformation of Consciousness in Ordeals of Initiation." *Ethos* 13 (2): 150–74.

Müller, Stephan, and Georg von Wangenheim. 2016. "Coevolution of Cooperation, Preferences, and Cooperative Signals in Social Dilemmas." SSRN Scholarly Paper ID 2526688. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=2526688.

Murray, Michael, and Lyn Moore. 2009. "Costly Signaling and the Origin of Religion." *Journal of Cognition and Culture* 9 (3): 225–45. https://doi.org/10.1163/156770909X12489459066264.

Nadler, Ronald D., and Eric S. Bartlett. 1997. "Penile Erection: A Reflection of Sexual Arousal and Arousability in Male Chimpanzees." *Physiology & Behavior* 61 (3): 425–32. https://doi.org/10.1016/S0031-9384(96)00454-4.

Ney, Alyssa. 2008. "Defining Physicalism." *Philosophy Compass* 3 (5): 1033–48. https://doi.org/10.1111/j.1747-9991.2008.00163.x.

Norenzayan, Ara. 2013. *Big Gods: How Religion Transformed Cooperation and Conflict*. 1 edition. Princeton: Princeton University Press.

———. 2014. “Does Religion Make People Moral?” *Behaviour* 151 (2–3): 365–84. https://doi.org/10.1163/1568539X-00003139.

Norenzayan, Ara, Azim F. Shariff, Will M. Gervais, Aiyana K. Willard, Rita A. McNamara, Edward Slingerland, and Joseph Henrich. 2016a. “The Cultural Evolution of Prosocial Religions.” *Behavioral and Brain Sciences* 39. https://doi.org/10.1017/S0140525X14001356.

———. 2016b. “Parochial Prosocial Religions: Historical and Contemporary Evidence for a Cultural Evolutionary Process.” *Behavioral and Brain Sciences* 39 (January): e29. https://doi.org/10.1017/S0140525X15000655.

Northover, Stefanie B., and Adam B. Cohen. 2017. “Understanding Religion from Cultural and Biological Perspectives.” In *The Handbook of Culture and Biology*, edited by José M. Causadias, Eva H. Telzer, and Nancy A. Gonzales, 55–77. John Wiley & Sons, Inc. https://doi.org/10.1002/9781119181361.ch3.

Nowak, Martin A. 2006. *Evolutionary Dynamics: Exploring the Equations of Life*. First Edition edition. Cambridge, Mass: Belknap Press.

Nowak, Martin A., Corina E. Tarnita, and Edward O. Wilson. 2010. “The Evolution of Eusociality.” *Nature* 466 (7310): 1057–62. https://doi.org/10.1038/nature09205.

O’Connor, Cailin. 2016. “The Evolution of Guilt: A Model-Based Approach.” *Philosophy of Science* 83 (5): 897–908. https://doi.org/10.1086/687873.

Ohtsubo, Yohsuke, and Esuka Watanabe. 2009. “Do Sincere Apologies Need to Be Costly? Test of a Costly Signaling Model of Apology.” *Evolution and Human Behavior* 30 (2): 114–23. https://doi.org/10.1016/j.evolhumbehav.2008.09.004.

Okasha, Samir. 2003. “Could Religion Be a Group-Level Adaptation of Homo Sapiens?: Darwin’s Cathedral: Evolution, Religion and the Nature of SocietyDavid Sloan Wilson; University of Chicago Press, 2002, Pp. V+268, Price $25 Hardback, ISBN 0-226-90134-3.” *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 34 (4): 699–705. https://doi.org/10.1016/j.shpsc.2003.09.009.

———. 2006. *Evolution and the Levels of Selection*. Oxford: Clarendon Press.

Owren, Michael J., Drew Rendall, and Michael J. Ryan. 2010. “Redefining Animal Signaling: Influence versus Information in Communication.” *Biology & Philosophy* 25 (5): 755–80. https://doi.org/10.1007/s10539-010-9224-4.

Pape, Robert. 2006. *Dying to Win: The Strategic Logic of Suicide Terrorism*. Reprint. Random House Trade Paperbacks.

Parke, Emily C. 2014. “Experiments, Simulations, and Epistemic Privilege.” *Philosophy of Science* 81 (4): 516–36. https://doi.org/10.1086/677956.

Patterson, Nick, Daniel J. Richter, Sante Gnerre, Eric S. Lander, and David Reich. 2006. “Genetic Evidence for Complex Speciation of Humans and Chimpanzees.” *Nature* 441 (7097): 1103–8. https://doi.org/10.1038/nature04789.

Pawlowitsch, Christina. 2008. “Why Evolution Does Not Always Lead to an Optimal Signaling System.” *Games and Economic Behavior* 63 (1): 203–26. https://doi.org/10.1016/j.geb.2007.08.009.

Peters, Joris, Klaus Schmidt, Oliver Dietrich, and Nadja Pöllath. 2014. “Göbekli Tepe: Agriculture and Domestication.” In *Encyclopedia of Global Archaeology*, edited by Claire Smith, 3065–68. Springer New York. http://link.springer.com.virtual.anu.edu.au/referenceworkentry/10.1007/978-1-4419-0465-2_2226.

Pettitt, Paul. 2015. “Landscapes of the Dead: The Evolution of Human Mortuary Activity from Body to Place in Pleistocene Europe.” In *Settlement, Society and Cognition in Human Evolution. .*, edited by Fiona Coward, Robert Hosfield, Matt Pope, and Francis Wenban-Smith, 258–73. Cambridge: Cambridge University Press.

Piff, Paul K., Daniel M. Stancato, Stéphane Côté, Rodolfo Mendoza-Denton, and Dacher Keltner. 2012. “Higher Social Class Predicts Increased Unethical Behavior.” *Proceedings of the National Academy of Sciences* 109 (11): 4086–91. https://doi.org/10.1073/pnas.1118373109.

Planer, Ronald J. 2017. “How Language Couldn’t Have Evolved: A Critical Examination of Berwick and Chomsky’s Theory of Language Evolution.” *Biology & Philosophy* 32 (6): 779–96. https://doi.org/10.1007/s10539-017-9606-y.

Pomiankowski, Andrew, and Yoh Iwasa. 1998. “Handicap Signaling: Loud and True?” Edited by Amotz Zahavi and Avishag Zahavi. *Evolution* 52 (3): 928–32. https://doi.org/10.2307/2411290.

Pounder, D. J. 1983. “RituaL Mutilation. Subincision of the Penis among Australian Aborigines.” *The American Journal of Forensic Medicine and Pathology* 4 (3): 227–29.

Poundstone, William. 1993. *Prisoner’s Dilemma: John von Neumann, Game Theory, and the Puzzle of the Bomb*. First Edition Thus edition. New York: Anchor.

Powell, Russell, and Steve Clarke. 2012. “Religion as an Evolutionary Byproduct: A Critique of the Standard Model.” *The British Journal for the Philosophy of Science* 63 (3): 457–86. https://doi.org/10.1093/bjps/axr035.

Power, Eleanor 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.

Powers, Simon T., Carel P. van Schaik, and Laurent Lehmann. 2016. “How Institutions Shaped the Last Major Evolutionary Transition to Large-Scale Human Societies.” *Phil. Trans. R. Soc. B* 371 (1687): 20150098. https://doi.org/10.1098/rstb.2015.0098.

Purzycki, Benjamin Grant, Joseph Henrich, Coren Apicella, Quentin D. Atkinson, Adam Baimel, Emma Cohen, Rita Anne McNamara, Aiyana K. Willard, Dimitris Xygalatas, and Ara Norenzayan. 2018. “The Evolution of Religion and Morality: A Synthesis of Ethnographic and Experimental Evidence from Eight Societies.” *Religion, Brain & Behavior* 8 (2): 101–32. https://doi.org/10.1080/2153599X.2016.1267027.

Rabett, Ryan J. 2018. “The Success of Failed Homo Sapiens Dispersals out of Africa and into Asia.” *Nature Ecology & Evolution* 2 (2): 212. https://doi.org/10.1038/s41559-017-0436-8.

Ramsey, Grant, and Andreas De Block. 2017. “Is Cultural Fitness Hopelessly Confused?” *The British Journal for the Philosophy of Science* 68 (2): 305–28. https://doi.org/10.1093/bjps/axv047.

Rappaport, Roy A. 1999. *Ritual and Religion in the Making of Humanity*. Cambridge Studies in Social and Cultural Anthropology 110. Cambridge, U.K. ; New York: Cambridge University Press.

Richerson, Peter, Ryan Baldini, Adrian V. Bell, Kathryn Demps, Karl Frost, Vicken Hillis, Sarah Mathew, et al. 2016. “Cultural Group Selection Plays an Essential Role in Explaining Human Cooperation: A Sketch of the Evidence.” *Behavioral and Brain Sciences* 39. https://doi.org/10.1017/S0140525X1400106X.

Richerson, Peter J., and Robert Boyd. 2005. *Not by Genes Alone: How Culture Transformed Human Evolution*. Chicago: University of Chicago Press.

Rigdon, Mary, Keiko Ishii, Motoki Watabe, and Shinobu Kitayama. 2009. “Minimal Social Cues in the Dictator Game.” *Journal of Economic Psychology* 30 (3): 358–67. https://doi.org/10.1016/j.joep.2009.02.002.

Riley, John G. 2001. “Silver Signals: Twenty-Five Years of Screening and Signaling.” *Journal of Economic Literature* 39 (2): 432–78. https://doi.org/10.1257/jel.39.2.432.

Riolo, Rick L., Michael D. Cohen, and Robert Axelrod. 2001. “Evolution of Cooperation without Reciprocity.” *Nature* 414 (6862): 441–43. https://doi.org/10.1038/35106555.

———. 2002. “Behavioural Evolution (Communication Arising): Does Similarity Breed Cooperation?” *Nature* 418 (6897): 500. https://doi.org/10.1038/418500a.

Rivers, Andrew M., and Jeff Sherman. 2018. “Experimental Design and the Reliability of Priming Effects: Reconsidering the ‘Train Wreck,’” January. https://doi.org/10.31234/osf.io/r7pd3.

Roberts, Gilbert, and Thomas N. Sherratt. 2002. “Behavioural Evolution (Communication Arising): Does Similarity Breed Cooperation?” *Nature* 418 (6897): 499–500. https://doi.org/10.1038/418499b.

Rosenstock, Sarita, and Cailin O’Connor. 2018. “When It’s Good to Feel Bad: An Evolutionary Model of Guilt and Apology.” *Frontiers in Robotics and AI* 5. https://doi.org/10.3389/frobt.2018.00009.

Ruxton, G. D., and H. M. Schaefer. 2011. “Resolving Current Disagreements and Ambiguities in the Terminology of Animal Communication.” *Journal of Evolutionary Biology* 24 (12): 2574–85. https://doi.org/10.1111/j.1420-9101.2011.02386.x.

Sandholm, William H. 2011. *Population Games and Evolutionary Dynamics*. Economic Learning and Social Evolution. Cambridge, Mass: MIT Press.

Sarkissian, Hagop. 2016. “Aspects of Folk Morality.” In *A Companion to Experimental Philosophy*, edited by Justin Sytsma and Wesley Buckwalter, 212–24. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118661666.ch14.

Scheuring, István. 2010. “Coevolution of Honest Signaling and Cooperative Norms by Cultural Group Selection.” *Biosystems* 101 (2): 79–87. https://doi.org/10.1016/j.biosystems.2010.04.009.

Schloss, Jeffrey P., and Michael J. Murray. 2011. “Evolutionary Accounts of Belief in Supernatural Punishment: A Critical Review.” *Religion, Brain & Behavior* 1 (1): 46–99. https://doi.org/10.1080/2153599X.2011.558707.

Schmidt, Klaus. 2000. “Göbekli Tepe, Southeastern Turkey: A Preliminary Report on the 1995-1999 Excavations.” *Paléorient* 26 (1): 45–54.

———. 2010. “Göbekli Tepe–the Stone Age Sanctuaries. New Results of Ongoing Excavations with a Special Focus on Sculptures and High Reliefs.” *Documenta Praehistorica* 37: 239–256.

Scott, James C. 2017. *Against the Grain: A Deep History of the Earliest States*. Yale University Press.

Scott-Phillips, T. C. 2008. “Defining Biological Communication.” *Journal of Evolutionary Biology* 21 (2): 387–95. https://doi.org/10.1111/j.1420-9101.2007.01497.x.

Searcy, William A., and Stephen Nowicki. 2005. *The Evolution of Animal Communication: Reliability and Deception in Signaling Systems*. Princeton, UNITED STATES: Princeton University Press. http://ebookcentral.proquest.com/lib/anu/detail.action?docID=485769.

Seyfarth, Robert M., and Dorothy L. Cheney. 2003. “Signalers and Receivers in Animal Communication.” *Annual Review of Psychology* 54 (1): 145–73. https://doi.org/10.1146/annurev.psych.54.101601.145121.

Shannon, C. E. 1948. “A Mathematical Theory of Communication.” *The Bell System Technical Journal* 27 (3): 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

Shariff, Azim F., and Ara Norenzayan. 2007. “God Is Watching You Priming God Concepts Increases Prosocial Behavior in an Anonymous Economic Game.” *Psychological Science* 18 (9): 803–809.

Shariff, Azim F., Aiyana K. Willard, Teresa Andersen, and Ara Norenzayan. 2016. “Religious Priming: A Meta-Analysis With a Focus on Prosociality.” *Personality and Social Psychology Review* 20 (1): 27–48. https://doi.org/10.1177/1088868314568811.

Shaver, J. H., and J. A. Bulbulia. 2016. “Signaling Theory and Religion.” In *Mental Religion*, edited by Jeffrey J. Kripal and Niki K. Clements, 101–17. Macmillan Interdisciplinary Handbooks.

Shaver, John H., Gloria Fraser, and Joseph A. Bulbulia. 2016. “Charismatic Signaling: How Religion Stabilizes Cooperation and Entrenches Inequality.” In *The Oxford Handbook of Evolutionary Psychology and Religion*, edited by Todd K. Shackelford and James R. Liddle. Oxford University Press. http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199397747.001.0001/oxfordhb-9780199397747-e-17.

Skoglund, Pontus, and Iain Mathieson. 2018. “Ancient Genomics of Modern Humans: The First Decade.” *Annual Review of Genomics and Human Genetics* 19 (1): 381–404. https://doi.org/10.1146/annurev-genom-083117-021749.

Skyrms, Brian. 2002. “Signals, Evolution and the Explanatory Power of Transient Information.” *Philosophy of Science* 69 (3): 407–28. https://doi.org/10.1086/342451.

———. 2003. *The Stag Hunt and the Evolution of Social Structure*. Cambridge: Cambridge University Press. http://public.eblib.com/EBLPublic/PublicView.do?ptiID=256680.

———. 2010. *Signals: Evolution, Learning, & Information*. Oxford; New York: Oxford University Press.

Smart, Ninian. 1989. *The World’s Religions : Old Traditions and Modern Transformations*. Cambridge : Cambridge University Press. https://trove.nla.gov.au/work/19032031.

Sober, Elliott, and David Sloan Wilson. 1999. *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Cambridge, Mass.: Harvard University Press.

Sosis, Richard. 2000. “Religion and Intragroup Cooperation: Preliminary Results of a Comparative Analysis of Utopian Communities , Religion and Intragroup Cooperation: Preliminary Results of a Comparative Analysis of Utopian Communities.” *Cross-Cultural Research* 34 (1): 70–87. https://doi.org/10.1177/106939710003400105.

———. 2003. “Why Aren’t We All Hutterites?” *Human Nature* 14 (2): 91–127. https://doi.org/10.1007/s12110-003-1000-6.

———. 2004. “The Adaptive Value of Religious Ritual: Rituals Promote Group Cohesion by Requiring Members to Engage in Behavior That Is Too Costly to Fake.” *American Scientist* 92 (2): 166–72.

———. 2005. “Does Religion Promote Trust?: The Role of Signaling, Reputation, and Punishment.” *Interdisciplinary Journal of Research on Religion* 1: 1–30.

———. 2006. “Religious Behaviors, Badges, and Bans: Signaling Theory and the Evolution of Religion.” In *Where God and Science Meet: How Brain and Evolutionary Studies Alter Our Understanding of Religion*, edited by Patrick McNamara, 1:61–86. Psychology, Religion, and Spirituality. Westport, Conn: Praeger Publishers.

———. 2009. “The Adaptationist-Byproduct Debate on the Evolution of Religion: Five Misunderstandings of the Adaptationist Program.” *Journal of Cognition and Culture* 9 (3): 315–32. https://doi.org/10.1163/156770909X12518536414411.

Sosis, Richard, and Candace Alcorta. 2003. “Signaling, Solidarity, and the Sacred: The Evolution of Religious Behavior.” *Evolutionary Anthropology: Issues, News, and Reviews* 12 (6): 264–74. https://doi.org/10.1002/evan.10120.

Sosis, Richard, and Eric R. Bressler. 2003. “Cooperation and Commune Longevity: A Test of the Costly Signaling Theory of Religion.” *Cross-Cultural Research* 37 (2): 211–39. https://doi.org/10.1177/1069397103037002003.

Sosis, Richard, and Jordan Kiper. 2014. “Religion Is More than Belief: What Evolutionary Theories of Religion Tell Us about Religious Commitments.” In *Challenges to Moral and Religious Belief: Disagreement and Evolution*, edited by Michael Bergmann and Patrick Kain. Oxford University Press. http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199669776.001.0001/acprof-9780199669776-chapter-14.

Sosis, Richard, and Bradley J. Ruffle. 2003. “Religious Ritual and Cooperation: Testing for a Relationship on Israeli Religious and Secular Kibbutzim.” *Current Anthropology* 44 (5): 713–22. https://doi.org/10.1086/379260.

Sparks, Adam, and Pat Barclay. 2015. “No Effect on Condemnation of Short or Long Exposure to Eye Images.” *Letters on Evolutionary Behavioral Science* 6 (2): 13–16. https://doi.org/10.5178/lebs.2015.35.

Spence, Michael. 1973. “Job Market Signaling.” *The Quarterly Journal of Economics* 87 (3): 355–374.

———. 2002. “Signaling in Retrospect and the Informational Structure of Markets.” *American Economic Review* 92 (3): 434–59. https://doi.org/10.1257/00028280260136200.

Sperber, Dan. 1996. *Explaining Culture: A Naturalistic Approach*. Oxford, UK ; Cambridge, Mass: Blackwell.

———. 2000. “An Objection to the Memetic Approach to Culture.” In *Darwinizing Culture: The Status of Memetics as a Science*, 163–173. http://dan.sperber.fr/wp-content/uploads/2009/09/meme.pdf.

Sripada, Chandra Sekhar, and Stephen Stich. 2006. “A Framework for the Psychology of Norms.” In *The Innate Mind Volume 2: Culture and Cognition.*, 280–301. Evolution and Cognition. New York, NY, US: Oxford University Press.

Stackhouse, Max L. 2007. *God and Globalization: Volume 4: Globalization and Grace*. A&C Black.

Stanford, P. Kyle. 2018. “The Difference between Ice Cream and Nazis: Moral Externalization and the Evolution of Human Cooperation.” *Behavioral and Brain Sciences* 41. https://doi.org/10.1017/S0140525X17001911.

Stark, Rodney. 1996. *The Rise of Christianity: A Sociologist Reconsiders History*. Princeton University Press.

Sterelny, Kim. 2006. “Memes Revisited.” *The British Journal for the Philosophy of Science* 57 (1): 145–65. https://doi.org/10.1093/bjps/axi157.

———. 2007. “SNAFUS: An Evolutionary Perspective.” *Biological Theory* 2 (3): 317–28. https://doi.org/10.1162/biot.2007.2.3.317.

———. 2012a. *The Evolved Apprentice: How Evolution Made Humans Unique*. Cambridge, Mass.: MIT Press.

———. 2012b. “A Glass Half-Full: Brian Skyrms’s Signals.” *Economics and Philosophy* 28 (01): 73–86. https://doi.org/10.1017/S0266267112000120.

———. 2016. “Cooperation, Culture, and Conflict.” *The British Journal for the Philosophy of Science* 67 (1): 31–58. https://doi.org/10.1093/bjps/axu024.

———. 2017a. “Cultural Evolution in California and Paris.” *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 62 (April): 42–50. https://doi.org/10.1016/j.shpsc.2016.12.005.

———. 2017b. “Religion Re-Explained.” *Religion, Brain & Behavior* 0 (0): 1–20. https://doi.org/10.1080/2153599X.2017.1323779.

Stich, Stephen. 2016. “Why There Might Not Be an Evolutionary Explanation for Psychological Altruism.” *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 56 (April): 3–6. https://doi.org/10.1016/j.shpsc.2015.10.005.

———. 2018. “Do We Really Externalize or Objectivize Moral Demands?” *Behavioral and Brain Sciences* 41. https://doi.org/10.1017/S0140525X18000183.

Stiner, Mary C. 2017. “Love and Death in the Stone Age: What Constitutes First Evidence of Mortuary Treatment of the Human Body?” *Biological Theory* 12 (4): 248–61. https://doi.org/10.1007/s13752-017-0275-5.

Stoljar, Daniel. 2010. *Physicalism*. New Problems of Philosophy. London ; New York: Routledge.

Számadó, Szabolcs. 2011. “The Cost of Honesty and the Fallacy of the Handicap Principle.” *Animal Behaviour* 81 (1): 3–10. https://doi.org/10.1016/j.anbehav.2010.08.022.

Szathmáry, Eörs, and John Maynard Smith. 1995. “The Major Evolutionary Transitions.” *Nature* 374 (6519): 227–32. https://doi.org/10.1038/374227a0.

Tomasello, Michael, and Amrisha Vaish. 2013. “Origins of Human Cooperation and Morality.” *Annual Review of Psychology* 64 (1): 231–55. https://doi.org/10.1146/annurev-psych-113011-143812.

Trivers, Robert L. 1971. “The Evolution of Reciprocal Altruism.” *The Quarterly Review of Biology* 46 (1): 35–57. https://doi.org/10.1086/406755.

Tucker, A. W. 1983. “The Mathematics of Tucker: A Sampler.” *The Two-Year College Mathematics Journal* 14 (3): 228–32. https://doi.org/10.2307/3027092.

Tylor, Edward Burnett. 1871. *Primitive Culture: Researches Into the Development of Mythology, Philosophy, Religion, Art, and Custom*. J. Murray.

Val, Aurore. 2016. “Deliberate Body Disposal by Hominins in the Dinaledi Chamber, Cradle of Humankind, South Africa?” *Journal of Human Evolution* 96 (July): 145–48. https://doi.org/10.1016/j.jhevol.2016.02.004.

Valen, L. van. 1973. “A New Evolutionary Law.” *Evolutionary Theory* 1: 1–30.

Van Elk, Michiel, Dora Matzke, Quentin Gronau, Maime Guang, Joachim Vandekerckhove, and Eric-Jan Wagenmakers. 2015. “Meta-Analyses Are No Substitute for Registered Replications: A Skeptical Perspective on Religious Priming.” *Frontiers in Psychology* 6. https://doi.org/10.3389/fpsyg.2015.01365.

Van Lawick-Goodall, Jane. 1968. “The Behaviour of Free-Living Chimpanzees in the Gombe Stream Reserve.” *Animal Behaviour Monographs* 1 (January): 161-IN12. https://doi.org/10.1016/S0066-1856(68)80003-2.

Wade, Lizzie. 2015. “Birth of the Moralizing Gods.” *Science* 349 (6251): 918–22. https://doi.org/10.1126/science.349.6251.918.

Wakeley, John. 2008. “Complex Speciation of Humans and Chimpanzees.” *Nature* 452 (7184): E3–4. https://doi.org/10.1038/nature06805.

Warner, Gregory. 2013. “How One Kenyan Tribe Produces The World’s Best Runners.” *All Things Considered*. NPR. http://www.npr.org/sections/parallels/2013/11/01/241895965/how-one-kenyan-tribe-produces-the-worlds-best-runners.

Watson-Jones, Rachel E., and Cristine H. Legare. 2016. “The Social Functions of Group Rituals.” *Current Directions in Psychological Science* 25 (1): 42–46. https://doi.org/10.1177/0963721415618486.

Watts, Ian, Michael Chazan, and Jayne Wilkins. 2016. “Early Evidence for Brilliant Ritualized Display: Specularite Use in the Northern Cape (South Africa) between ∼500 and ∼300 Ka.” *Current Anthropology* 57 (3): 287–310. https://doi.org/10.1086/686484.

Watts, Joseph, Simon J. Greenhill, Quentin D. Atkinson, Thomas E. Currie, Joseph Bulbulia, and Russell D. Gray. 2015. “Broad Supernatural Punishment but Not Moralizing High Gods Precede the Evolution of Political Complexity in Austronesia.” *Proceedings of the Royal Society of London B: Biological Sciences* 282 (1804): 20142556. https://doi.org/10.1098/rspb.2014.2556.

Watts, Joseph, Oliver Sheehan, Quentin D. Atkinson, Joseph Bulbulia, and Russell D. Gray. 2016. “Ritual Human Sacrifice Promoted and Sustained the Evolution of Stratified Societies.” *Nature* advance online publication (April). https://doi.org/10.1038/nature17159.

Watts, Joseph, Oliver Sheehan, Simon J. Greenhill, Stephanie Gomes-Ng, Quentin D. Atkinson, Joseph Bulbulia, and Russell D. Gray. 2015. “Pulotu: Database of Austronesian Supernatural Beliefs and Practices.” *PLOS ONE* 10 (9): e0136783. https://doi.org/10.1371/journal.pone.0136783.

Weaver, Ryan J., Rebecca E. Koch, and Geoffrey E. Hill. 2017. “What Maintains Signal Honesty in Animal Colour Displays Used in Mate Choice?” *Phil. Trans. R. Soc. B* 372 (1724): 20160343. https://doi.org/10.1098/rstb.2016.0343.

Weisberg, Michael. 2006. “Robustness Analysis.” *Philosophy of Science* 73 (5): 730–42. https://doi.org/10.1086/518628.

———. 2013. *Simulation and Similarity: Using Models to Understand the World*. New York: Oxford University Press.

West, Sandra A., Ashleigh S. Griffin, and Andy Gardner. 2007. “Social Semantics: Altruism, Cooperation, Mutualism, Strong Reciprocity and Group Selection.” *Journal of Evolutionary Biology* 20 (2): 415–32. https://doi.org/10.1111/j.1420-9101.2006.01258.x.

Whitehouse, Harvey. 1995. *Inside the Cult: Religious Innovation and Transmission in Papua New Guinea*. Oxford : New York: Clarendon Press.

———. 2018a. “Dying for the Group: Towards a General Theory of Extreme Self-Sacrifice.” *Behavioral and Brain Sciences* 41. https://doi.org/10.1017/S0140525X18000249.

———. 2018b. “Dying for the Group: Towards a General Theory of Extreme Self-Sacrifice.” *Behavioral and Brain Sciences*, February, 1–64. https://doi.org/10.1017/S0140525X18000249.

Whitehouse, Harvey, and Jonathan A. Lanman. 2014. “The Ties That Bind Us: Ritual, Fusion, and Identification.” *Current Anthropology* 55 (6): 674–95. https://doi.org/10.1086/678698.

Wilks, Matti, Emma Collier-Baker, and Mark Nielsen. 2015. “Preschool Children Favor Copying a Successful Individual over an Unsuccessful Group.” *Developmental Science* 18 (6): 1014–24. https://doi.org/10.1111/desc.12274.

Willard, Aiyana K, Azim F Shariff, and Ara Norenzayan. 2016. “Religious Priming as a Research Tool for Studying Religion: Evidentiary Value, Current Issues, and Future Directions.” *Current Opinion in Psychology*, Social priming, 12 (December): 71–75. https://doi.org/10.1016/j.copsyc.2016.06.003.

Williams, G. C. 1966. *Adaptation and Natural Selection*. Princeton: Princeton University Press.

Wilson, David Sloan. 2003. *Darwin’s Cathedral: Evolution, Religion, and the Nature of Society*. 1 edition. Chicago,Ill.: University Of Chicago Press.

Winegard, Bo M., Tania Reynolds, Roy F. Baumeister, Benjamin Winegard, and Jon K. Maner. 2014. “Grief Functions as an Honest Indicator of Commitment.” *Personality and Social Psychology Review* 18 (2): 168–86. https://doi.org/10.1177/1088868314521016.

Xygalatas, Dimitris. 2013. “Effects of Religious Setting on Cooperative Behavior: A Case Study from Mauritius.” *Religion, Brain & Behavior* 3 (2): 91–102. https://doi.org/10.1080/2153599X.2012.724547.

———. 2018. “What Fuses Sports Fans?” *Behavioral and Brain Sciences* 41. https://doi.org/10.1017/S0140525X18001814.

Xygalatas, Dimitris, Panagiotis Mitkidis, Ronald Fischer, Paul Reddish, Joshua Skewes, Armin W. Geertz, Andreas Roepstorff, and Joseph Bulbulia. 2013. “Extreme Rituals Promote Prosociality.” *Psychological Science* 24 (8): 1602–5. https://doi.org/10.1177/0956797612472910.

Zahavi, Amotz. 1975. “Mate Selection—A Selection for a Handicap.” *Journal of Theoretical Biology* 53 (1): 205–14. https://doi.org/10.1016/0022-5193(75)90111-3.

———. 1977. “The Cost of Honesty: Further Remarks on the Handicap Principle.” *Journal of Theoretical Biology* 67 (3): 603–605.

Zahavi, Amotz, and Avishag Zahavi. 1997. *The Handicap Principle: a Missing Piece of Darwin’s Puzzle*. New York: Oxford University Press.

Zahedzadeh, Giti. 2017. “Designed to Fail: Modeling Terrorism’s Losing Battle.” *Journal of Terrorism Research* 8 (2). https://doi.org/10.15664/jtr.1272.

Zollman, Kevin J. S. 2013. “Finding Alternatives to Handicap Theory.” *Biological Theory* 8 (2): 127–32. https://doi.org/10.1007/s13752-013-0107-1.

Zollman, Kevin J. S., Carl T. Bergstrom, and Simon M. Huttegger. 2013. “Between Cheap and Costly Signals: The Evolution of Partially Honest Communication.” *Proceedings of the Royal Society B: Biological Sciences* 280 (1750). https://doi.org/10.1098/rspb.2012.1878.