UNIVERSITY OF CALIFORNIA,
IRVINE

The Coevolution of Commitment and Cooperation over Human History

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Philosophy

by

Saira Khan

Dissertation Committee:
Professor P. Kyle Stanford, Chair
Distinguished Professor Brian Skyrms
Chancellor's Professor Simon Huttegger
Chancellor's Professor Jeffrey A. Barrett
Professor Kim Sterelny

2023

TABLE OF CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

Page

ACKNOWLEDGEMENTS

I would like to express the deepest appreciation to my committee chair, Professor P. Kyle Stanford and my co-advisor, Professor Kim Sterelny. Without their guidance and enthusiasm for my ideas, this dissertation would not have been possible. I am grateful for their time, their academic rigor, and the challenges they pushed me to overcome.

I would also like to thank my other committee members, Distinguished Professor Brian Skyrms and Chancellor's Professors Simon Huttegger and Jeffrey A. Barrett, who aided in my intellectual development throughout graduate school. These professors have guided me through a vast array of fields of study with utmost proficiency.

Finally, I would like to thank my partner, Josiah Lopez-Wild, who held my hand from start to finish, who bolstered my confidence when I was doubtful, and who valiantly withstood my countless rants about the coevolution of commitment and cooperation over human history. Thanks also to my dog, Johnny, who did none of the above but is a good boy.

VITA

Saira Khan

2015 B.A. (Hons) in Philosophy, Politics and Economics, University of Oxford

2018 PG Cert in Central Banking and Financial Regulation, University of Warwick

2023 M.A. in Philosophy, University of California, Irvine

2023 Ph.D. in Philosophy, University of California, Irvine

FIELD OF STUDY

Logic and Philosophy of Science

ABSTRACT OF THE DISSERTATION

The Coevolution of Commitment and Cooperation over Human History

by

Saira Khan

Doctor of Philosophy in Philosophy

University of California, Irvine, 2023

Professor P. Kyle Stanford, Chair

In this dissertation, I argue that a significant but previously overlooked factor in the evolution of human prosociality is the coevolution of commitment and cooperation. Commitments can serve to secure mutually beneficial interaction in the face of short-term incentives to cheat. They do so by simultaneously incentivising an agent to act in line with her signalled intent and acting as a marker through which trustworthy partners can be identified. I show how different methods of undertaking commitments in our evolutionary history have enabled more sophisticated forms of cooperation over time which, in turn, create the selective environment for the evolution of increasingly effective commitments.

I argue that pre-linguistic commitments secured the stability of cooperation in early hominin group hunting as early as two million years ago and that hunting laid the foundations for larger-scale cooperation. In particular, I argue that signalling participation in group hunting opens up opportunities for beneficial interaction by way of creating a social bond. This potential for future beneficial interaction increases the cost of defection from acting against one's signalled intent since defection may entail exclusion from these benefits. This incentivises cooperation, acting as

a commitment. Though I use hunting as the illustrative example, there are likely many other cooperative activities in our ancestral environment that possess the same formal features of commitment, for example, participation in joint gathering or collection of resources. This small-scale collaboration via commitment then enabled the formation of our proto-language and the expansion of the group into a multi-level society. This created a selective environment for the emergence of more sophisticated communication among members of the wider group, culminating in language.

Language enabled us to undertake commitments via explicit and implicit promising. Explicit promises take the form of direct assertions of one's intentions to cooperate. Implicit promises often take the form of gossip about a third party. To illustrate, signalling one's disapproval of a normatively-laden behaviour can lead another to believe that the disapproving agent will not act in like manner, and if she does, she may be subject to exclusion from the benefits of current or future interaction. Analogously for signalling one's approval. I suggest these two forms of linguistic commitment allowed for more effective communication in cooperative ventures and expanded the range of cooperative activity that could be undertaken. Language allows us to commit with more specificity, to commit to spatially and temporally remote events and to make conditional commitments, among other benefits. Language and reputation sharing also increases opportunities for detection and punishment of false commitments.

A further development in our evolutionary history – the capacity to view some norms as universally applicable or *externalised* – adds even more power and scope to our commitment practices. To directly assert one's attitudes toward an externalised, usually moral, norm or gossip

about another's violation of an externalised norm is to suggest one's own commitment to that norm. This is because such norms are considered applicable to all, including oneself. Implicit commitments which take the form of gossip concerning cultural norm-violation only apply where the interlocutor has reason to believe that the gossiper is part of the same cultural group and occupies the same social role as the person about whom they are gossiping. However, commitments based on externalised norms require fewer inferences on the part of the interlocutor. She need only believe that the agent takes the norm to be universally applicable. Furthermore, in expressing attitudes towards externalised norms, the agent not only advertises her future behaviour but simultaneously reveals to her interlocutor that she is willing to exclude others who do not behave similarly, since she takes this norm to be applicable to all. This secures better correlated interaction.

Finally, the cooperation enabled by earlier forms of commitment permitted the development of institutionalised third-party punishment, which offered a new enforcement mechanism for commitments. That is, we see the development of commitments backed by physical or financial consequences. This enforcement is carried out by third parties and it is typically conducted in an organised manner with the backing of common resources. Modern examples include legal contracts or oaths. Yet again, these commitments allow for safer cooperation among an even wider network of individuals. Third-party enforcement strengthens the credibility of commitments and can, as a result, lower the cognitive demands of trust. It also provides new partner detection and control mechanisms. So, I argue that as the range of our cooperative enterprises extend so, too, do our means of safely engaging in cooperation through the use of more effective forms of commitment.

Alongside highlighting an important feature in the evolution of human prosociality – the coevolution of commitment and cooperation – this dissertation explains why we commit in the various ways that we do. That is, this dissertation sheds light on how we commit without language, why we believe linguistic promises, why we take other people's gossip statements to be commitments and why we have codified commitments in our institutions. Each of these is a response to a changing cooperative environment in which it is increasingly important to find effective ways to ensure that we make choices that serve our long-run interests and which signal our trustworthiness to others.

Chapter 1: Introduction

"He who was ready to sacrifice his life, as many a savage has been, rather than betray his comrades, would often leave no offspring to inherit his noble nature."

Charles Darwin, The Descent of Man (1871)

1.1 What is cooperation?

Cooperative interaction between two or more individuals can lead to synergies which could not have been achieved by agents in isolation. For example, a group of lionesses can kill a deer when one alone may not be able to shepherd the target away from her herd. Social insects achieve synergies through building protective structures to benefit the whole group. Division of labour, seen frequently between the sexes in tribal communities, also allows for collective profits as time can be spent equally on different tasks and expertise in one skill is easier to build than expertise in many. Cooperation can also manage risk through the reciprocal sharing of resources within a group when one agent may be incapacitated by injury or illness. Humans cooperate in a number of different ways and cooperation is often spontaneous, even proactive, between individuals who have never interacted before. Yet despite being mutually beneficial, cooperation may not be stable. If profit is differently distributed, or if coordinating action is noisy and involves costly mistakes, some agents may do better by not cooperating or may choose to avoid risks even if they would.

A key question for philosophers and biologists is how cooperative behaviour could have arisen when such behaviour may entail a cost to one's reproductive fitness, yet natural selection favours those traits which contribute to reproductive success. In other words, why would an individual

carry out a cooperative behaviour which benefits another if it is apparently costly to perform? There are many existing explanations, including kin selection, group selection, reciprocal altruism and indirect reciprocity, among others. My contribution concerns the role of commitment in the evolution of cooperation. In particular, I argue a significant but previously overlooked factor in the evolution of human prosociality is the coevolution of commitment and cooperation over our history.

To speak of the coevolution of commitment and cooperation, we must first be clear on what we mean by "cooperation". Hamilton (1964) famously classified social behaviour into four types based on the fitness consequences for the actor and recipient in an interaction, as shown in Table 1. Whether a behaviour is beneficial or costly depends on the lifetime fitness consequences of the behaviour and the absolute fitness effect.[1]

Effect on actor	+	-
+	Mutual benefit	Selfishness
-	Altruism	Spite

Table 1: Hamiltonian classification of social behaviours.

[1] This table is adapted from The Genetic Evolution of Social Behavior (1964). Hamilton himself termed "altruism" and "selfishness", and the term "spite" appears in his later work (Hamilton 1970). We call the +/+ cell mutual benefit. See Forber and Smead (2015) for a criticism of Hamilton's social classification. It will not matter much to my account whether this classification is robust since I merely use it for conceptual clarity when referring to the extant literature on the evolution of cooperation to situate the commitment account among other theories. In what follows, I will primarily refer to *interactions* or *outcomes* in which both agents benefit as situations of "mutually benefit", rather than referring to the *behaviours* or *actions* of a particular agent as mutually beneficial as in Hamilton's schema. This is because, if an action is straightforwardly beneficial to an agent, it does not require explanation – the behaviour or trait which contributes to such behaviour would be selected for.

While many existing accounts of the evolution of cooperation focus on explaining altruism – behaviours which are costly to the actor – cooperation also includes those behaviours which are mutually beneficial. Indeed, in the ordinary sense of the word, we often take cooperation to refer to collaborative action to achieve a joint goal. This is manifested in early humans in activities such as group hunting and defence, division of labour, and collective care of offspring. In modern humans, cooperation may take the form of trade, the advent of agricultural economies and state organisation. The terms "altruism" and "cooperation" have been used differently in different parts of the social evolution literature (West et al. 2006, 2007). I follow the work of Sachs et al. (2004) in using cooperation to refer to those social behaviours which provide a benefit to the recipient but may be *either* beneficial or costly to the agent.[2] This covers cases of both altruism and mutual benefit according to Hamilton's schema.

Theories of the evolution of cooperation can be broadly classified into those which appeal to direct fitness benefits and those which appeal to indirect fitness benefits. *Direct* benefits are proposed to explain *mutually beneficial* cooperation while *indirect* benefits explain *altruistic* cooperation.[3] Direct benefit explanations appeal to the fact that, even though the actor undergoes a cost whilst benefitting another, this cost is outweighed by a (usually downstream) benefit to the actor herself. An example is cleaner fish who clean the mouths of larger fish, who in turn refrain from eating them.[4] Theories of the evolution of cooperation that appeal to direct fitness benefit often involve

[2] West et al. (2007) argue that this definition of cooperation is too permissive since it includes as cooperation those actions which are only accidentally beneficial to the actor. As an example, they cite an elephant producing dung: this is beneficial to the elephant as it empties waste and is also beneficial to the dung beetle. However, to call "cooperation" those acts which provide a benefit as a byproduct of the behaviour seems amiss. The authors instead prefer that "cooperation" refers to behaviours which are selected for *because of* their beneficial impact on the recipient. This may well be a more precise definition, but it will not matter much for our purposes.

[3] See also West et al. (2007) on this point.

[4] One might alternatively say that direct benefit explanations use repeated games to convert a one-shot altruism problem into mutual payoff maximisation in the long run.

a shared interest in cooperation and a mechanism which makes non-cooperation costly. This can be seen in the theories of reciprocal altruism[5] (Trivers 1971), punishment in repeated games (Rapoport & Chammah 1965; Axelrod & Hamilton 1981; Axelrod 1984), strong reciprocity (Bowles & Gintis 2004) or indirect reciprocity (Alexander 1987; Nowak & Sigmund 1998, 2005; Pollock & Dugatkin 1992; Mohtashemi & Mui 2003). Since these models see cooperative behaviour as to the benefit of the individual enacting it, they are instances of cooperation as mutual benefit. More will be said on the details of these theories and how they interact with the account to be offered in this dissertation later in this chapter.

Indirect fitness benefits refer to the inclusive fitness gains an individual may achieve through their impact on the reproductive success of related individuals, as well as through the impact on their own reproductive success. An example is the allomothering seen in vervet monkeys, where related females such as older sisters or grandmothers often care for young, according to their relatedness. Indirect benefit explanations of the evolution of cooperation explain cooperative behaviour by way of kin discrimination (Hamilton 1963), limited dispersal of genetically related individuals, or cooperation preferentially directed to other cooperators who are not genetically related to the individual (Robson 1990; Skyrms 1996). In the latter case, cooperative behaviours indirectly benefit the agent through promoting the shared "cooperative gene". Since these behaviours are costly to the individual agent but beneficial to related individuals or to the cooperative gene, they are instances of cooperation as altruism.

[5] Though termed "altruism" this theory explains mutually beneficial behaviour since it appeals to the long-term benefit of the cooperative agent. Indeed, Hamilton (1996: 263) also notes this theory was misnamed.

The literature is complicated further by the introduction of the term "weak altruism" in group selection. This is used in Sober and Wilson's (1998) prominent treatment of the evolution of altruism but differs from the Hamiltonian classification. Wilson (1975, 1977) defines a behaviour as weakly altruistic if it leads to a decrease in the fitness of the individual relative to the rest of the group. However, such behaviours can be selected for because they increase an individual's *direct fitness* if the individual derives a benefit from the group's success. As such, weakly altruistic behaviours can also be mutually beneficial. This is especially true if we consider that groups with greater resources may experience increased reproductive success. As such, group selection (Wilson 1975, 1977; Colwell 1981; Wilson & Colwell 1981; Wilson & Sober 1994; Sober & Wilson 1998) and cultural group selection (Richerson & Boyd 1998) explanations of the evolution of cooperation generally cross-cut the direct/indirect explanatory divide. To clear up the conceptual landscape, we will appeal to the following distinctions made by West and colleagues (2007: 422).

Term	Effect on actor	Effect on recipient
Weak altruism	- or +	+
Altruism	-	+
Mutual benefit	+	+
Cooperation	- or +	+

Table 2: Glossary of terms.

To summarise, cooperation includes both instances of mutual benefit and of altruism. Explanations of the evolution of cooperation which appeal to direct benefit are those which explain cooperation as mutual benefit. Explanations which appeal to indirect benefit are those which explain cooperation as altruism. The account to be offered in this dissertation is one of the coevolution of commitment and cooperation. Since it will appeal to direct benefit, it is an account of the evolution

of cooperation as mutual benefit rather than cooperation as genuine altruism. Of course, there are many interactions in which it is difficult to determine the extent to which behaviour impacts direct or indirect fitness, and therefore the extent to which it is mutually beneficial or altruistic. For example, geographic dispersal may be altruistic if it reduces local competition for resources but mutually beneficial if it involves the agents moving to better habitats for themselves (West et al. 2007). However, as many of my examples will concern interactions where the payoffs of cooperation are immediate, it will be easy to see that these are situations of mutual benefit.

The cases of interest to us will be those in which a particular outcome is mutually beneficial but it is profitable for an agent to free-ride or cheat. In these cases, we would expect the actor to do so as this secures better relative fitness consequences. For example, in provision of a public good, although mutual benefit can be gained from contribution, an individual does better by not contributing and instead free-riding off the contributions of others. In dyadic interactions, if an actor can secure the benefit in question by deception, it would not make sense for her to cooperate fully and pay the cost of cooperation. Alternatively, the behaviour which contributes to mutual benefit might be risky. Finally, mutually beneficial outcomes may only be realisable with a certain level of trust which strong assumptions of rationality will prevent. These problems will be explored in the coming chapters. We will see that commitment is one way in which mutually beneficial cooperation can be secured in the face of short-term temptations to cheat. First, we must survey the existing literature on commitment and introduce the concept of commitment I will be working with.

1.2 Existing treatments of commitment

Here, I discuss the literature on commitment in game theory. The game theory literature considers many different forms of commitment, including but not limited to: threats and promises; conditional versus unconditional commitments; and intentional versus unintentional commitments. I will address these distinctions in the discussion to follow. In the next section, I highlight those features of commitment that I believe are salient to explaining its role in the evolution of cooperation, drawing upon some of the ideas presented here.

Schelling (1960) introduced commitments in his work on bargaining and conflict in game theory. Under Schelling's formulation, commitments operate by irreversibly reducing an agent's payoffs for a particular action or by removing options for the agent, such that it induces the other agent to choose in her favour. The alteration of the agent's incentives could be achieved by worsening one's payoff in the event of non-fulfilment or by delegation of control to another. Such commitments are communicated (verbally or non-verbally) to the other agent, which alters her expectations. In Schelling's words, "the commitment is a strategic move, a move that induces the other player to choose in one's favor. It constrains the other player's choice by affecting his expectations… a rational second player can be constrained by his knowledge that the first player has altered his own incentive structure" (1960: 122-3). It is supposed that these are situations in which, by undertaking a commitment, an agent optimises her long-term utility even though the alternative option looks suboptimal in the short-term.[6] For example, an agent who tears off her steering wheel in a game

[6] Of course, an agent can also commit to choices which would not be better for her in the long run and can commit where she does not have an incentive to cheat in the short-term, but these will not be the cases of interest for explaining the evolution of cooperation.

of chicken will ultimately win the game of chicken by inducing her partner to swerve. Though tearing off one's steering wheel may appear costly in the short-term, it serves the agent's long-term interest in winning the game of chicken.

For Schelling, commitments are credible if they change payoffs such that acting in line with one's commitment is now optimal. For example, in the following Chicken payoff matrix, it is necessary that Row's threat imposes a cost or disutility of 2 on the option of swerving, rendering Drive the dominant option, as below.[7] Since Drive is the dominant option for Row, Column expects Row to drive, securing the outcome, (R2,C1), in which Row wins. Nash equilibria are represented in green and changes to payoffs as a result of commitment are indicated in red. Note that representing the commitment via changed payoffs in the 2x2 game changes the game itself. However, payoffs are usually understood to be fixed in games. Schelling also shows how commitment strategies can be represented in a larger game matrix which captures reduced payoffs without altering the game itself. However, for simplicity of exposition, let us continue with Schelling's 2x2 presentation.

	C1 (Swerve)	C2 (Drive)
R1 (Swerve)	3,3	2,4
R2 (Drive)	4,2	1,1

Table 3: Chicken or Hawk-Dove.

[7] In reality, commitments may be credible if they make an option more *likely* to be chosen, rather than making sure the option is chosen with certainty. Here, however, I present Schelling's original dominance criteria.

	C1 (Swerve)	C2 (Drive)
R1 (Swerve)	1,3	0,4
R2 (Drive)	4,2	1,1

Table 4: Chicken or Hawk-Dove with commitment to R2.

Here, the commitment takes the form of a threat which secures a win for one agent at the expense of the other. The intuitive difference between threats and promises is that promises make an option more attractive for the agent's partner while threats make an option less attractive. Yet, in game theory, this amounts to the same thing since it is relative payoffs which matter. So, though I use promises as examples in this dissertation, the same analysis will apply for threats which secure mutually beneficial outcomes.

Indeed, Hirshleifer (2003) simply characterises commitments as either preemptive or reactive. Commitments, on his account, do not change the payoffs in the game. Instead, they change the order of play. This characterisation blurs the distinction between threats and promises since a reactive commitment in a Prisoner's Dilemma (see Table 5) such as "if Row chooses R1, I will respond with C1; if Row chooses R2 I will respond with R2" contains elements of both a conditional promise and a threat. Reactive commitments are made by players who have the second execution move and can condition their choice on their partner's. Preemptive commitments allow the player to take the first execution move in the game. Intuitively, this corresponds to a standard promise where an agent's action does not depend on her partner's.

Since, for Hirshleifer, commitments determine order of play, he argues there is no role for commitment in games with no first-mover advantage, for example in coordination games such as

the Stag Hunt (Table 6). Schelling, however, does see commitment as valuable in these contexts. Following Schelling, I hold there is a role for commitment in any simultaneous move game where mutual gains can be achieved. Later in this chapter, we will see how commitment operates with respect to the Stag Hunt. First, we will discuss more of the nuances in the notion of commitment.

	C1	C2
R1	3,3	1,4
R2	4,1	2,2

Table 5: Prisoner's Dilemma.

	C1 (Stag)	C2 (Hare)
R1 (Stag)	3,3	0,2
R2 (Hare)	2,0	1,1

Table 6: Stag Hunt.

What are the means of ensuring a commitment is credible? Nesse (2003) distinguished four kinds of enforcement mechanisms for commitment which confer credibility to the signal. Self-enforcing commitments are secured by the fact that they make an option impossible, for example, tearing off one's steering wheel or burning a bridge behind oneself. Contractual commitments are secured by external incentives controlled by third parties. Here, the option to renege on the commitment is still available to the agent but no longer advantageous. An example is enforcement via a lease agreement, where breaking the lease entails a financial cost. The final two types of commitments – emotional and reputational – are what Nesse refers to as "subjective" commitments. Emotional commitments are enforced by feelings such as pride and guilt. Reputational commitments are

backed by the pledge of one's reputation, for example, in taking a public oath. Of course, many commitments cross-cut these enforcement mechanisms. Marriage is an example of a commitment enforced by a legal contract, reputational pledges and emotions.[8]

Frank (1988; 2011) has argued that the evolution of cooperation is at least partially dependent on the realisation of promises backed by subjective enforcement mechanisms, particularly emotion. Such commitments serve to change the optimal option of the agent by adding internal disutility or utility to their downstream subjective payoffs.[9] For example, the continuity of a partnership is better guaranteed with an emotional commitment because this renders the option to pursue new opportunities for partnering *undesirable* to the agent. While it may not be optimal in the short-term for an agent to forego an opportunity to engage in an illicit affair, virtuousness is in her long-term interest in future interactions. Emotions such as love render the subjective satisfaction of loyalty to her current partner sufficiently high that it is now optimal for the agent to forego the affair and remain loyal (Frank 1988). This virtuousness is signalled to other agents, changing their expectations of behaviour and allowing other agents to identify the virtuous agent as trustworthy for future interaction, securing material gains for the virtuous agent in the long run.[10]

Schelling (2003) notes commitments may also be made credible by involuntary features of a person, for example, personal traits, qualities, abilities and techniques (constraints as well as capacities). Borrowing an example from Conrad's *The Secret Agent,* Schelling writes of the secret

[8] Note, commitment signals may not be perfectly reliable and the threshold for a signal to induce cooperation may depend on the number of trustworthy agents in the population. It may also be altered by the costs of scrutiny, dispositions to detect and punish false signalling, and access to further information. See Frank (1988) for a model of signal reliability. These issues will be discussed in Chapters 2-5.

[9] This is in reference to the proximate causes of commitment behaviour rather than their ultimate causes. See Mayr (1961) and Tinbergen (1968).

[10] See also Hirshleifer (1984) on emotions as a guarantors to commitments.

agent that "rigging the nitroglycerin in his clothing is deliberate; being the kind of person who would actually detonate the nitroglycerin is not" (Schelling 2003: 53). Schelling further notes that devotion to a deity or culture of honor which obligates fidelity does not automatically make one committed – an agent becomes committed only if he makes an oath – but being a person who is devoted in such a way is to be the *kind* of person who can swear credibly.

This discussion is pertinent to cultural group membership and norms. To change one's payoffs, there must be an enforcement mechanism by which one's incentives can be recognisably altered, and group membership and norms provide a cultural common ground wherein such commitments can be communicated and understood. They may make one the *kind* of person who can swear credibly. In addition to providing a cultural common ground for recognition of subjective enforcement mechanisms, groups may also impose external enforcement mechanisms such as fines or imprisonment. This will become important in our discussion of institutionalised commitments in Chapter 5. Yet many cases of commitment to be discussed in this dissertation do not rely on such external enforcement. Instead, the credibility of the commitment is ensured only by the *intrinsic value of future cooperation*. That is, most commitments important for the evolution of human cooperation are reputationally-backed. One who commits and does not follow through risks foregoing fitness-enhancing opportunities for future interaction with others.

Finally, it is important to note that commitments might not have arisen solely for the purpose of cooperative signalling. In the first instance, the signal would not have been recognised by the receiver as a mechanism for reliable partner choice.[11] Frank suggests that emotional commitment

[11] Modelling work presented in the Appendix to the dissertation addresses how signals could come to be understood as commitments.

mechanisms initially served as "self-control devices" (Frank 2011: 72). That is, they reduce the temptation for short-term gains in pursuit of long-term ones. Luce and Raiffa have suggested such a mechanism is at play when a person goes on a diet. They write, "he announces his intention, or accepts a wager that he will not break his diet, so that later he will *not* be free to change his mind and to optimize his actions according to his tastes at *that* time" (Luce & Raiffa 1957: 75). Others, such as Elster (2000) and Gibbard (1990) have also suggested that agents will do better by pre-committing to a long-term goal, and that natural selection would have supported the development of such a "normative control system" (Gibbard 1990: 56-7). This idea has been developed further in the work of McClennen (1990) on resolute choice in diachronic choice theory.

To recap, the most prominent treatments of commitment in the game-theoretical literature appear in the works of Schelling (1960; 2003), Frank (1988; 2011), and Hirshleifer (1984; 2003). In these accounts, commitments are broadly moves which an agent takes to ensure her long-term interest, even though taking such a move may appear sub-optimal in the short run. More recent game-theoretical treatments have operationalised commitment differently, in terms of requests for aid or subjective utility derived from repeated interaction with the same individual (Han 2013; Back 2007). We have seen that commitments may take the form of either threats or promises. Commitments can be made credible by a number of features; self-enforcing, contractual, emotional, and reputational mechanisms. Commitments can even be made credible by involuntary features of a person. In the next section, we tease out what features of commitment will be important for the present thesis.

1.3 Defining commitment

Though there is no one canonical definition of commitment, the traditional accounts of commitment presented in the previous section agree on the following: a commitment is a move which an agent takes at time *t* that increases the probability[12] she carries through an option *X* at time *t+n* (Schelling 1960; Frank 1988, 2011).[13] This is both a necessary and sufficient condition to something being a commitment. Indeed, despite the differences in the accounts of Hirshleifer, Han and Back, each author sees themselves as building on the work of Schelling (Hirshleifer 2003: 77; Han 2013: 120; Back 2007: 6). As such, we will work with the more traditional account. The commitment works by changing the incentives of the committing agent through adding either physical, contractual, emotional, reputational, or other costs to choosing against one's commitment, incentivising option *X* at time *t+n*. If our criteria for the requisite payoff change is dominance and the agent is perfectly rational, the commitment ensures she carries through option *X* with certainty. In the typical contexts in which we see commitment deployed, option *X* is the option which is optimal in the long run though it may not be attractive to the agent in the short-term. Let us call this the wide-scope definition of commitment.

However, the definition we will need for this dissertation is narrower. In particular, we are not concerned with commitments that are not in service of social interactions. Indeed, under the previous definition of commitment, accepting a wager that one will not break one's diet counts as a commitment but this has no relevance to the evolution of cooperation. So let us employ the

[12] The "probability" here does not to refer to the agent's credence in her own actions as this is too cognitive a notion of commitment. By saying that a commitment "increases the probability" of her so acting, I simply mean that following through on option *X* is more likely given the commitment.
[13] See also Sterelny (2012).

following narrow-scope definition, which applies to social interaction and supplements the above notion with concepts from signalling game theory (Lewis 1969).

> **Definition:** *A commitment is a pre-play signal in a strategic interaction taken at time t, that increases the sender's relative payoff for carrying through option X at time t+n, and increases the receiver's probability*[14] *of the sender carrying through option X.*[15]

This is a more precise codification of the mechanism at play in Schelling's work and amends the wide-scope definition for application to a strategic interaction. A strategic interaction is a situation in which the payoffs of one agent are dependent upon the choices of another. By "relative payoff" I mean that the commitment either makes option *X* more attractive than another option, or it makes the other option less attractive than option *X.* The payoff change for the sender is actual, though this in turn changes the agent's expected payoff, too. For the receiver, there is a change in her expected payoff without a change in her actual payoffs, as only her beliefs about the sender's actions have changed. Note that this definition does not invoke intentional states, so the sender need not believe she has made a commitment in order for her signal to be a commitment.[16]

We can commit to something which we would be independently motivated to do, for example, caring for a neighbour's garden. In this case, the commitment would add further reason to care for the garden since it changes our relative payoffs for carrying through this option. However, these

[14] We assume that the receiver's credences are in some way correctly connected to the sender's motivation to act. That is, her belief that that the sender is more likely to take option *X* is correlated with the sender's likelihood of in fact taking option *X.*

[15] Note that the conditional commitment need not be formulated differently. That is, we do not need an additional clause specifying "option *X* if opponent choses option *Y*". This is because the commitment itself induces the other player to choose in the committed player's interest. It thus serves to induce *Y.*

[16] See, for example, Gilbert (2013).

kinds of commitments will not typically capture the scenarios of interest for this dissertation. I am concerned with explaining the evolution of cooperation where there is incentive to defect. Here, the commitment can motivate the agent to do something which she would not otherwise do. Hirshleifer (1984) suggests this is what distinguishes commitments from mere forecasts. Similarly, Schelling writes, "if I have every reason to sue for damages in the event you do me harm, saying so is not a threat, it is merely a communication, what I call a warning. Arranging incentives so that I must sue whether I want to or not and communicating that I must do so is the threat" (Schelling 2003: 52). Though I do not believe we have to limit the definition of commitment to refer to only those signals which motivate an agent to do what she would not otherwise do, the sorts of commitments which are important for explaining the evolution of cooperation in light of the pull of defection are indeed those which motivate an agent to do what she would not otherwise do.

As we will see in the following chapters, sometimes, but not always, the alterations of the sender's relative payoffs will come *by way of* the change in receiver expectations.[17] It is because the receiver expects the sender to act in a particular way that she will exclude the sender if she acts otherwise.

[17] Mischkowski et al. (2018) show that the expectations of the promisee's partner incentivises promise-keeping (Charness & Dufwenberg 2006) as well as the duty-bound feelings of the promisee that incentivises promise-keeping (Vanberg 2008). Furthermore, they find an interaction effect: "a promisor might be particularly concerned about not disappointing a promisee's expectations that were caused by his promise, such that the effect of promising is to make the promisor more sensitive to the promisee's expectations than he would have been had he not made the promise" (Mischkowski et al. 2018: 4). This mechanism echoes the operation of commitments I have provided here – commitments involve both a change in sender motivations and receiver expectations. They find evidence of this interaction effect though vignette studies. Subjects are asked to imagine that they are the prospective buyer of a product from a seller who is out of town. Some have been told that they promised the seller they would buy their product upon their return, while others were told they had told the seller that they merely planned to do so, explicitly stating they were not making a promise. The participants are also told the degree of belief the seller has that the participant will buy the product, ranging from completely certain the participant will not, to completely certain she will. Subjects are asked how likely it is that they will buy the product, despite finding it elsewhere at a lower price. It is found that an increase in the seller's expectations cause a greater increase in subjects' reported likelihood of buying from seller in the promise conditions than in the no promise conditions. Specifically, in the promise condition, when the seller's expectation of performance was 0 per cent, 37 per cent of participants honoured the promise, compared to 67 per cent when the seller's expectation of performance was 100 per cent. In the no promise condition, these numbers were 10 per cent and 30 per cent respectively.

This is when commitments are reputationally enforced. Other times, the change in the sender's relative payoff is what changes the receiver expectations. This is the case when the sender undertakes a legal contract. The change in her payoffs here is not dependent on the change in receiver expectations since the commitment is enforced by a third party. Furthermore, in some cases, option *X* will only become available to an agent when they have already undergone the commitment. Consider, for example, signing a lease agreement. It is not possible to act in accord or disaccord with the commitment to pay rent until one has signed the agreement. Nonetheless, this does not alter the fact that the important feature of a commitment is that it alters downstream payoffs for the sender (there is cost to not paying rent) and changes the receiver's probability of the sender carrying through this option.

The characterisation of a commitment as a "pre-play signal" implies that the commitment is designed to transmit information. This is in contrast to a "cue" which is something which can be used to infer an agent's future behaviour, though it did not evolve for a communicative purpose. For example, the fearful cry of a monkey in response to the sight of a predator is a cue that conveys information about the predator's whereabouts to its conspecifics. However, in this case, where the cry is produced by a reflexive fear response, the mechanism of production is unaffected by the consequences of the sign's use by the receiver (Godfrey-Smith 2014). In contrast, signals are deployed by a sender for use in state-to-act or act-to-act coordination. Where cues convey information only as a by-product of an activity, "signals have been shaped by natural selection for the specific purpose of conveying information and thereby influencing others' behavior, ultimately impacting both the signaler's and the recipient's fitness" (Laidre & Johnstone 2013).

There is a separate issue of whether a commitment signal must be voluntary. There are some obvious cases where non-commitment signalling is involuntary. Consider, for example, the sub-personal signalling of a hormones within the body or the sexual swelling and colouration of a female baboon's genitals during periods of their menstrual cycle is a signal. Both have evolved for the purpose of communicating receptiveness to reproduction. However, neither are voluntary. For the purpose of this dissertation, I am interested in voluntary signals. This is because I am concerned with commitments where agents have the option to do otherwise. A signal is voluntary if an agent can intervene in order to send the signal or not. As such, signals which are based on immutable characteristics of the agent will not be of interest in this dissertation. For example, features of an agent such as physical strength or group membership may convey an agent's ability to successfully engage in a group hunt, but these will not constitute commitments as I am conceiving of them, since commitments are moves in a strategic game.[18] As noted earlier in our discussion of Schelling (2003), these features of an agent may make his commitment more *credible,* but the agent is not committed until he voluntarily signals.

It is also important to see that the difference between a cue and a signal does not concern the difference between an intentional and an unintentional act. A sender can send a signal which is misinterpreted by the receiver – the meaning of the signal is not as intended. The signal has still evolved to convey information, but the information gleaned by the receiver may be otherwise than its evolved purpose. For example, if Sandy gossips about the behaviour of a third party, Betty may take this to convey information about Sandy's future behaviour. However, gossip has evolved as a reputation sharing device, so the original purpose of the signal was to share information about

[18] We might instead refer to this as the agent having an "obligation" or duty". Recall, the definition of commitment requires that the agent has issued a pre-play signal which she had the option not to send.

the third party rather than to advertise one's future behaviour. The utterance may nonetheless constitute a *cue* about Sandy's future behaviour, but it is only a signal insofar it has evolved to communicate such information. Here, the aspect of the signal that pertains to Sandy's future behaviour is unintentional. In this way, we can see that the sender of a signal can make commitments without knowing so. That is, she can send signals which are intended to transmit information about another but nonetheless change the relative payoffs of her future actions.

As we will see in Chapters 2 and 3, something which initially was only a cue might also evolve to be a signal. Barrett and Skyrms (2017) discuss how a cue can turn into a signal through a process of positive feedback and "ritualization".[19] For instance, an animal may naturally bare its teeth before biting and this is a cue to its future behaviour, conveying information to another animal. Yet, just as a receiver might evolve to exploit the fixed dispositions of a sender, it is also possible that the sender evolves to act in a way that exploits the receiver's pre-dispositions to respond to the sender's cues. Barrett and Skyrms (2017) term this a "sensory-manipulation game" which, along with a "cue-reading game" constitute a signalling game in which agents coevolve responses to one another in a way that is advantageous to both. During "ritualization", the sender's disposition to bare its teeth on the basis of the receiver backing away is positively reinforced, and can thereby come to mean a threat signal rather than simply a cue. We will see how cues such as showing up to a hunting excursion and gossiping about the behaviours of third parties will come to constitute commitments to one's own future actions in Chapters 2 and 3.

[19] See also Hauser (1997).

Of course, the characterisation of commitment given above will include various forms of commitment: threats, promises, conditional, unconditional, intentional, unintentional, preemptive, reactive and self-oriented commitments. Since I am concerned with the evolution of cooperation, I am concerned with commitments in strategic interactions rather than with commitments which apply to the individual. This account provided here focuses on promises, both conditional and unconditional, though it could equally well be extended to threats. Many of the commitments explored in this dissertation will also be intentional as in explicit promises and contractual agreements. Although, as will be seen in Chapter 3, implicit commitment can also be unintentional, since all that is required for a behaviour to count as a commitment is a change in the agent's payoffs, a signal which demonstrates this, and a change in their partner's expectations. Now that we have a precise definition of commitment, we can explicate how it operates to secure mutually beneficial outcomes in game thoery.

1.4 Operationalising commitment in classical game theory

In this section, I will present the costs of defecting on one's commitment as a reduced payoff in a one-shot game. This is how Schelling introduced the concept of commitment in game theory. However, in many real-life scenarios, the cost of defection comes from reduced opportunities for partnering in repeated games, which would be captured in an evolutionary game.[20] This is presented in the next section. First discussing the one-shot game will give us a clearer basis on which to understand the operation of commitment in an evolutionary context.

[20] As will be explored in Chapter 5, there can be physical and financial cost to reneging on one's commitment, too.

Another way we can understand the difference between a presentation of commitment in classical game theory and a presentation of commitment in evolutionary game theory is by invoking the distinction between proximate and ultimate causation (Mayr 1961; Tinbergen 1968). Proximate causes govern the responses of an organism to its environment – they generate behaviour via cognitive, social, psychological, physical or chemical processes – while ultimate causes explain the function of that behaviour in terms of its contribution to the organism's fitness. The disutility of defection in the one-shot game tracks proximate mechanisms for commitment behaviour – it is what directly motivates the agent to act in line with her signalled intent. The cost of defection in the case of evolutionary game theory tracks ultimate causation. At the level of ultimate causation, the increased cost of defection is captured in the fact that agents who commit and subsequently defect fare worse in the evolutionary dynamics since they are not chosen in repeated interactions, but this does not itself change the game being played. In most of the previous sections, I referred to the ultimate causes of commitment behaviour – the effect commitment has on opportunities for repeated interaction or potential punishment. However, reference was also made to proximate mechanisms such as love in Frank's (1988) account.

While both Schelling and Hirshleifer used normal form games in their presentations of commitment, Schelling himself notes that "while it is instructive and intellectually satisfying to see how such tactics as threats, commitments, and promises can be absorbed into an enlarged, abstract "supergame" (game in "normal form"), it should be emphasized that we cannot learn anything about those tactics by studying games that are already in normal form. The objects of our study, namely, these tactics together with the communication and enforcement structures that they depend on, and the timing of the moves, have all disappeared by the time the game is in normal

form." (Schelling 1960: 156). Much of this detail will be filled in in subsequent chapters. In a deviation from Schelling, I will represent commitment strategies in extensive form. This, of course, is still a simplification of real-world commitments which involve differing means of enforcement and signalling.

The Chicken game, though useful for illustrating what a commitment is, does not model the cooperative scenarios of interest for this dissertation. In particular, I am concerned with situations in which mutual benefit is to be gained from cooperation. Many of the situations are well represented by a Stag Hunt. This game captures situations in which mutual benefit is attainable but risky. In the Prisoner's Dilemma, an individual agent does best by defecting on another where, in the Stag Hunt, an individual does best by choosing in like manner to the other agent. Mutual cooperation in the Prisoner's Dilemma can achieve *global* optimality, but not *individual* optimality – it therefore captures situations where there is a conflict between individual rationality and mutual benefit, while such interests are aligned in the Stag Hunt. As such, the Stag Hunt is more apt for representing situations of collective action whilst the Prisoner's Dilemma may be better used to represent situations such as competition over resources. Of course, *repeated* Prisoner's Dilemmas capture scenarios where mutual gain is achievable over the course of multiple interactions, and for this, we will use a version of the evolutionary game presented in the next section.[21]

[21] For more on the relationship between Prisoner's Dilemma and Stag Hunt games in modelling the evolution of cooperation, see Skyrms (2003). It is shown that the shadow of the future in repeated interaction will serve to transform a Prisoner's Dilemma into a Stag Hunt. We could also use a Public Goods game or Threshold Public Goods game to represent cooperative interactions. However, this will only be useful for collective action problems involving multiple individuals and is difficult to represent in extensive form in the manner employed here.

In the Stag Hunt, players simultaneously choose to engage in a cooperative activity or not, and whether mutual gains are realised depends on whether both choose to cooperate without knowing what the other will choose. Both (R1,C1) and (R2,C2) are Nash equilibria. As such, we must provide an explanation of how mutually beneficial cooperation can arise in this game. We do so here with commitment. First, we must consider the payoffs that result from no commitment, commitment to Stag and commitment to Hare.

	C1 (Stag)	C2 (Hare)
R1 (Stag)	3,3	0,2
R2 (Hare)	2,0	1,1

Table 7: Stag Hunt.

	C1 (Stag)	C2 (Hare)
R1 (Stag)	3,3	0,2
R2 (Hare)	0,0	-1,1

Table 8: Stag Hunt with commitment to R1.

	C1 (Stag)	C2 (Hare)
R1 (Stag)	1,3	-2,2
R2 (Hare)	2,0	1,1

Table 9: Stag Hunt with commitment to R2.

The extensive form of the game is shown below. There is an information set over Row's choice to represent the fact that the game is simultaneous – Row does not know what Column chooses. Indeed, in collective hunting, although agents may embark on the hunt together, they do not know

whether the other will cooperate or defect when the prey arises. In the case of no commitment, Column cannot determine whether Row will play Stag or Hare as Row's best strategy is mixed. She prefers to play Stag if Column plays Stag but Hare if Column plays Hare. As a result, Column does not in effect choose between payoffs 3 or 1 but rather the expected payoff according to the mixed equilibrium strategy where Row plays Stag with probability 0.5 and Hare with probability 0.5. Therefore, on the middle branch, Column's expected payoff is 1.5.

However, if Row commits to Stag, by definition, her payoff for playing Hare is reduced. As such, Column expects that Row will play Stag no matter her choice. If this is so, she prefers to play Stag, yielding both payoffs of 3. Here, commitment takes the form of an unconditional promise, since moves are simultaneous and Row is not aware of what Column chooses. Commitment in the Prisoner's Dilemma would look much the same, but cooperation would be secured by a conditional threat to Defect, rather than a promise to Cooperate (since this promise alone would incentivise the other to Defect).[22]

[22] Hirshleifer (2003) believes a threat coupled with a conditional promise will serve to bolster credibility in the Prisoner's Dilemma, but we will ignore this as we are working with Schelling's (1960) notion of credibility which turns on a change in the sender's relative payoffs.

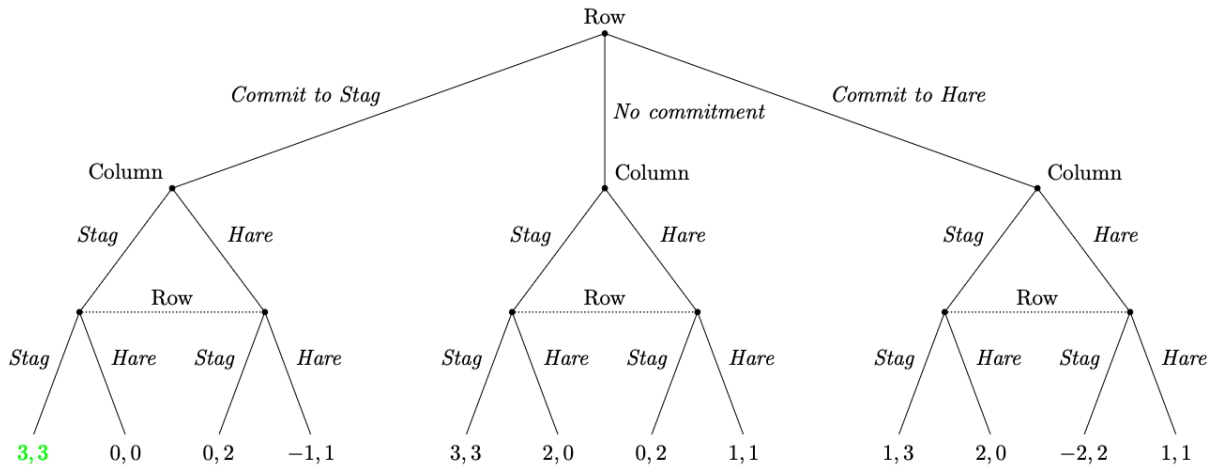

Figure 1: Commitment in extensive form Stag Hunt.

Now, let us consider the case where both players have the power to commit, as demonstrated in Table 10 and Figure 2 (CS, CH, and NC stand for Commit to Stag, Commit to Hare, and No Commitment, respectively).

		Commit to Stag		No commitment		Commit to Hare	
		S	H	S	H	S	H
Commit to Stag	S	3,3	0,0	3,3	0,2	3,1	0,2
	H	0,0	-1,-1	0,0	-1,1	0,-2	-1,1
No commitment	S	3,3	0,0	3,3	0,2	3,1	0,2
	H	2,0	1,-1	2,0	1,1	2,-2	1,1
Commit to Hare	S	1,3	-2,0	1,3	-2,2	1,1	-2,2
	H	2,0	1,-1	2,0	1,1	2,-2	1,1

Table 10: Payoff matrix for the Stag Hunt where both players can commit.

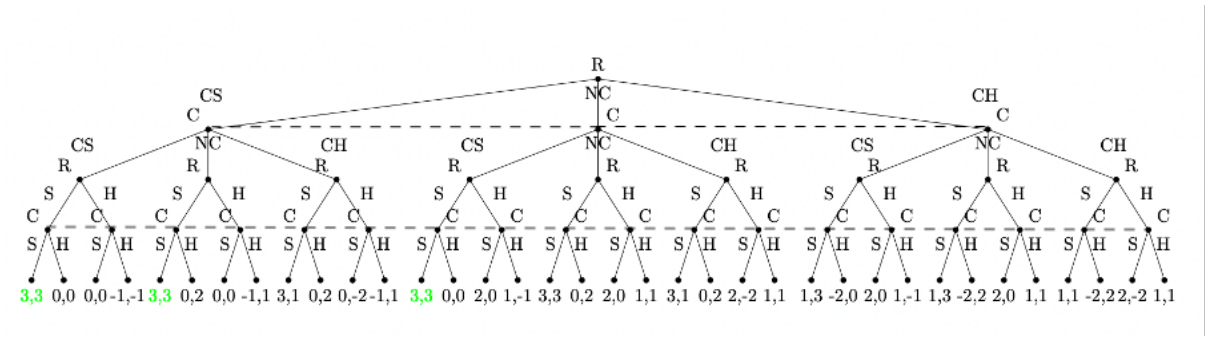

Figure 2: Commitment in extensive form Stag Hunt where both players can commit.

There is an information set over both Column's choice to play Stag or Hare and Column's choice to commit. This means that the commitments are simultaneous. The agent does not know whether Row has committed.[23] We solve this tree first by finding the Nash equilibria of each of the 9 subgames. This gives us the following reduced payoff matrix. We can see that there are three Nash equilibria, each where at least one player commits. Thus, even if both players have the power to commit, a unilateral commitment is sufficient to secure the best outcome in the Stag Hunt, since it effectively changes the receiver's expectations of cooperation.

	Commit to Stag	No commitment	Commit to Hare
Commit to Stag	3,3	3,3	0,2
No commitment	3,3	1.5,1.5	1,1
Commit to Hare	2,0	1,1	1,1

Table 11: Reduced payoff matrix for the extensive form Stag Hunt.

[23] The solution to sequential commitment involves at least one player committing to Stag.

Many cooperative scenarios also call for *conditional* promises. That is, a promise to reciprocate if a partner chooses to trust the other and engage in an interaction. This is captured in the extensive form Trust game (Berg et al. 1995). In the Trust game, the players are given an initial endowment, one player (the trustor) can choose to share a proportion of their endowment with a second player (the trustee). If resources are sent to the trustee, they are multiplied. The trustee then has the option to benefit the trustor in turn by resharing some or all of this increased sum. If agents are self-interested, the trustor expects that the trustee will keep the whole of the transferred resources so will choose to transfer nothing. However, this is not what is seen in experiments (Berg et al. 1995). Berg and colleagues suggest that trusting behaviour is rewarded as it is an instance of generosity. As we will see, commitment can also reliably secure the cooperative outcome. Below is the general form of a Trust game, where $0 < s < r < t$. We consider the simplest case where the resources are split equally between the agents upon cooperation.

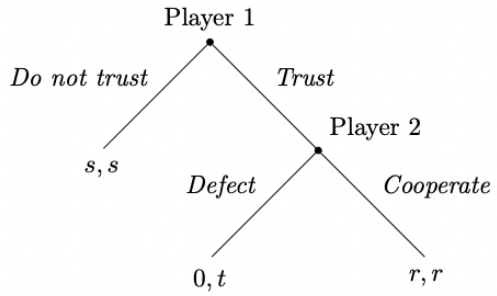

Figure 3: The Trust game.

Commitment in this game takes the form of a conditional promise on the part of Player 2. Player 2 promises to cooperate and share the multiplied endowment with Player 1 if Player 1 chooses to trust her and engage in the game rather than choosing the exit option. Many ordinary cooperative

interactions take this form: if an agent promises to help another with collecting resources but her potential partner would suffer from subsequent defection on the promise, the partner might choose not to engage in a cooperative activity with the agent. Here, a commitment may serve to change the optimal option of the committed agent such that defection is no longer profitable and her promise to cooperate on resource collection is now credible. This ensures the partner's safe engagement in the cooperative activity and the realisation of mutual gains. To represent the extended game with commitment, we must first present the change in payoffs that would make a commitment to cooperate credible.

Suppose our game involves payoffs $s = 5$, $r = 10$, $t = 20$. That is, the initial endowment is 5, and the multiplier is 4. If Player 2 credibly commits to cooperate, her defect payoff, t, must incur a disutility of 11 so $t' = 9$. We need not consider the case of commitment to defection since this is already the optimal option. The extended game is thus as follows.

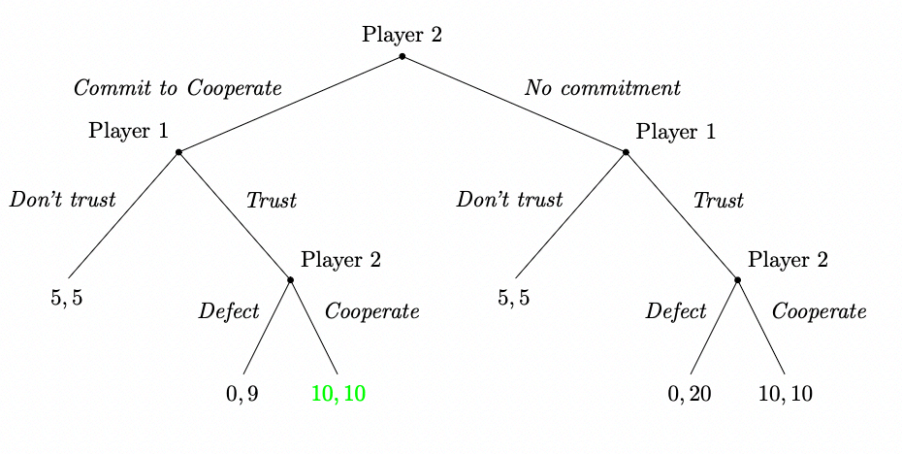

Figure 4: The Trust game with commitment.

By backward induction, the unique solution of the game is where Player 2 commits to cooperation, Player 1 trusts her, and Player 2 cooperates with a reward of (10,10). This is in contrast to the previous Nash equilibrium where Player 2 has an incentive to defect, which means that Player 1 does not trust her and a lower payoff (5,5) is attained for both players. This commitment is conditional in the sense that it depends on the other player's action, but the action here is not analogous to the action taken by the committing player. Instead, Player 1 faces the choice of playing the subsequent game or not.

In this section, I have illustrated how commitments can operate to secure mutually beneficial outcomes in one-shot cooperative interactions such as the Stag Hunt and Trust Game. I follow Schelling in conceptualising commitment as a change in the committed agent's relative payoffs and a change in her partner's expectations of her action, which induces the partner to choose in the committed agent's favour. Unlike the works of those before me, I have codified the increased disutility of acting against one's signalled intent as a reduced subjective payoff in the one-shot *extensive* form game. We may think of this as capturing the operation of the proximate mechanisms of commitment, such as the subjective disutility of reduced opportunities for repeated interaction or punishment – for example, fear of ostracism, shame or guilt. Now that we have understood how commitment operates more precisely in game theory, we can apply the concept in an evolutionary context.

1.5 Operationalising commitment in an evolutionary context

In the preceding analysis, the cost of defecting on one's commitment have been codified as a reduced payoff in a one-shot game, changing the game being played. We may think of this as

modelling the operation of the proximate mechanism of commitment – commitment changes the subjective payoff of acting in contradiction to one's signalled intent by way of guilt, fear or some other psychological phenomenon. However, at the level of ultimate causation, the increased cost of defection is largely constituted by reduced opportunities for partnering in repeated games. Indeed, the majority of commitments I discuss in this dissertation are enforced by one's reputation. Here, I will construct a simplified model to elucidate how reputationally-backed commitment secures cooperation in an evolutionary context and thereby elucidate how commitment operates at the level of ultimate causation. We will see in Chapter 5 that commitment may also include other non-reputational penalties for reneging, such as institutionally imposed fines or imprisonment. For now, let us ignore these additional features. First, we replace our payoffs with variables for generality as below, where $w > x > y > z$.

	C1 (Stag)	C2 (Hare)
R1 (Stag)	*w, w*	*z, x*
R2 (Hare)	*x, z*	*y, y*

Table 12: Generalised Stag Hunt.

The straightforward expected payoffs to Stag hunting and Hare hunting for Row are as follows. P_R denotes Row's credences and P_C denotes Column's credences.

$$EV(R1) = P_R(C1)w + P_R(C2)z$$

$$EV(R2) = P_R(C1)x + P_R(C2)y$$

However, pre-play commitment signalling introduces additional elements. First, the agents signal whether they will play Stag or Hare, then play the Stag Hunt. The person who commits and reneges becomes a less preferred partner for future interaction as long as her defection is detected. The agent who signals Stag and follows through on her commitment has better access to opportunities for future interaction. The payoffs therefore depend both on the signals sent in the pre-play game and subsequent action. In this spirit, let us say that the agent who reneges on a commitment undergoes some punishment cost which represents exclusion from the benefits of current and future interaction. Since this is captured in the payoff for reneging, we do not need to include an analogous parameter for inclusion in the benefits of current and future cooperation upon following through on a commitment. Of course, in a dynamical model, we would not need to capture the cost of exclusion from the benefits of current or future interaction in a single term at all, since this would play out in the dynamics. However, the following simplification suffices for the conceptual clarity we seek at present. Now, we have the following expected payoffs for Row.[24] Let *SR1* indicate that Row signals Stag and *SR2* indicate that Row signals Hare. Now, we have the following expected payoffs.

$$EV(SR1 \ \& \ R1) = P_R(C1)w + P_R(C2)z - P_R(SC1 \ \& \ C2)l$$

$$EV(SR1 \ \& \ R2) = P_R(C1)x + P_R(C2)y - P_R(d)c$$

Where $P_R(d)$ represents Row's expected probability of defection being detected, c represents the cost of punishment (most commonly, exclusion from the benefits of current or future interaction),

[24] Row has six total strategies*: SR1 & R1, SR1 & R2, SR2 & R1, SR2 & R2, ~S & R1,* and *~S & R2.* The payoff functions listed in-text are those that are important for our current discussion. That is, the payoff functions for those strategies where the agent commits to cooperation.

and *l* represents the cost of enacting punishment toward agents who have signalled cooperation and reneged. We will assume that punishment costs are the responsibility of cooperators, not defectors, and that social exclusion is a low-cost form of punishment so the value of *l* is negligible. This will be true in most cooperative interactions but is a more questionable assumption when we deal with public goods, where excluding someone might involve undergoing a cost to drive them away. This is discussed in more detail in Chapter 2. With this model in hand, we may speak more precisely about some of the ideas in my commitment framework.

Following Schelling (1960), my concept of commitment changes receiver expectations of the sender's action. Why is this so? The reason why $P_C(R1)$ increases as a result of Row's choice of *SR1* is because *SR1 & R2* entails a punishment cost, *c,* of exclusion from interaction. It is worth noting that, for the sender to be effectively deterred from playing *SR1 & R2*, it is necessary that *l* be sufficiently small such that the receiver's punishment is not sub-optimal for her. If *l* is sufficiently small, the pre-play signal incentivises *SR1 & R1.* In order to understand why $P_C(R1)$ increases, we would need to appeal to the evolutionary dynamics. When we relinquish the common knowledge assumptions of classical game theory, the receiver does not have access to the sender's fitness payoffs. However, suppose there are receivers whose $P_C(R1)$ reduces upon receipt of the signal *SR1.* These agents are incentivised to play Hare. Their payoffs are lower than those agents whose response to *SR1* is to increase their credence in $P_C(R1)$, since these agents are incentivised to play Stag and consequently receive a higher payoff. Thus, agents whose response to *SR1* is to reduce $P_C(R1)$ fare worse in the evolutionary dynamics. As such, on average, $P_C(R1)$ will increase as a result of Row's choice of *SR1.* Cooperators are thus able to identify and interact with other

cooperators on the basis of pre-play signalling. Agents who successfully correlate interaction receive the higher payoff *w*, rather than *z*.[25]

Second, what is trustworthiness in this model? Trustworthiness is simply the sender's likelihood of playing Stag. The receiver's perceived trustworthiness of the sender is captured in the receiver's expectation that the sender plays Stag, so $P_C(R1)$. We may contrast trustworthiness with honesty. Honesty is to signal and follow through on one's signalled intent. Agents who signal Hare and follow through on Hare are honest but are not trustworthy. Furthermore, there is no incentive to signal Hare since, if Row signals Hare, the equilibrium payoffs to the agents are (*y,y*). So Row does no better (in expectation) than if she had not signalled. As such, agents only signal Stag, if they signal anything at all. Perceived trustworthiness is based on both past plays and commitment. Once we have linguistic communication, perceived trustworthiness will also be bolstered by *news* of Row's previous plays as well as past observations. We assume that agents learn by reinforcement, such that yielding a positive payoff for *SR1 & R1* on previous turns makes it more likely that the agent chooses *SR1 & R1* in the future. That is, we can effectively learn about an agents trustworthiness from their past record of play.[26]

[25] This is assuming there is an exit or partnering option, such that cooperators are able to opt out of playing the Stag Hunt with a particular partner after the initial signalling stage and choose others preferentially on the basis of whether they have signalled commitment. This may not be an innocuous assumption in the case of small hominin groups, where there are few partnering opportunities. Nonetheless, evidence presented in Chapter 2 shows that joint activity even in the small-scale social worlds fit the commitment model.

[26] One can imagine scenarios in which this does not hold. For example, perhaps punishment is issued on a third "strike" for defection. If this is so, two instances of defection may increase one's credence that the agent will cooperate on the next round in order to avoid punishment. Alternatively, if an agent has accrued plenty of resources as a result of previous cooperation, this may lead her to defect in future activities, since cooperation is costly and the benefit is not worthwhile to her (though one might question whether this is defection at all if the agent does not stand to gain from it). However, these cases are comparatively rare when considered alongside all the instances in which past cooperation or defection *is* indicative of future behaviour. It therefore pays, on average, for a person to make such an inference. That past behaviour is indicative of future behaviour is an assumption made in many evolutionary models of the evolution of cooperation and is well-evidenced in the literature on partner choice. In economic games, participants preferentially choose to interact with, and give greater sums of money to, agents who were cooperative in a previous game (Barclay & Willer 2007; Sylwester & Roberts 2010).

Third, how does commitment improve access to opportunities for future interaction? If the receiver is incentivised to cooperate with the sender, the sender has generated opportunities for beneficial interaction should the agents meet again. Earlier, we spoke of $P_C(R1)$ increasing in subsequent rounds of the game. Commitment achieves this since the strategy *SR1 & R1* increases $P_C(R1)$, incentivising *C1* on the part of the receiver since this yields a higher payoff than *C2* if Row plays *R1*. As detailed earlier, $P_C(R1)$ increases as a result of pre-play commitment signalling combined with acting in accordance with one's signalled intent – that is, when Row chooses the strategy *SR1 & R1*.

Fourth, commitment changes the motivations of the sender and the information of the receiver. What does this look like in our model? The motivation-altering aspect of a signal is that, where $EV(R2) = P_R(C1)x + P_R(C2)y$, with pre-play signalling we have $EV(SR1\ \&\ R2) = P_R(C1)x + P_R(C2)y - P_R(d)c$. That is, pre-play signalling introduces a cost to reneging for the sender of the signal. The signal is information-carrying for the receiver because, as was noted earlier, the strategy *SR1 & R1* increases Column's $P_C(R1)$ since *SR1* in the pre-play stage is informative – it carries information about the sender's action when we reach the Stag Hunt stage, since the signal has introduced a cost to reneging, $P_R(d)c$. Notice also that there has been no change in the receiver's payoffs as she has not signalled. So her payoffs for Stag and Hare are $EV(C1) = P_C(R1)w + P_C(R2)z$ and $EV(C2) = P_C(R1)x + P_C(R2)y$, respectively. Her increased likelihood of cooperation comes by way of increasing $P_C(R1)$. This makes stag hunting more profitable than hare hunting due to a larger multiplier, but her payoff function has not changed.

Finally, we can now ask more precisely: under what parameters is cooperation profitable in this simplified model? We can see that it depends on the probability of Column's choosing to cooperate and play Stag, as well as on the relative cost of punishment and probability of defection being detected. In the ordinary Stag Hunt, cooperation, *R1*, is only profitable depending on the relative probabilities of Column's choices and the attached payoffs. However, with commitment signalling in advance, *SR1 & R1,* will be profitable as long as $P_R(d)c$ or c is sufficiently high. Indeed, in the evolutionary dynamics, the reason that commitment signalling is stable is that the agents who signal Stag and play Stag are preferentially chosen in future interactions and the agents who signal Stag and play Hare are not, so fare worse. Furthermore, since agents have access to information about the past plays of others with whom they have interacted, they can choose to interact with those agents for whom they believe the probability of playing Stag is high.[27]

Since we are concerned with the evolution of cooperative traits, I will say something on the mechanism of transmission. The transmission mechanism of commitment practices will largely be cultural, transmitted both horizontally and vertically through social learning. Specifically, signals of commitment might be socially acquired through observation and mimicking, direct teaching, or institutionalisation. Cultural transmission will be most important when environmental conditions are positively but imperfectly correlated across generations, so each generation will benefit from acquiring information that could not be transmitted genetically since it is not encoded in the germ line. This is what we expect to be the case in the changing cooperative environments described in this dissertation. However, it is possible that commitment practices are also genetically transmitted when environmental conditions were more stable. Indeed, some neuroscientific studies show a

[27] Of course, initial plays are risky since past information is unavailable. Commitment signalling is a useful tool in this context.

genetic basis for cooperative behaviour and find that damage to particular brain regions make people exhibit sociopathic behaviour (Gintis 2011). If this is so, commitment practices will have been established in the population via gene-culture coevolution. That is, through the cultural construction of society and its changing nature, we can provide new environments for fitness-enhancing genetic changes in the individual which have further impacts on cultural practices.

In this section, I have constructed a simplified model illustrating how, in an evolutionary context, a fitness cost to reneging on commitment will be realised over multiple interactions. This fitness cost takes the form of exclusion from cooperative interaction. Note that it is not the purpose of the dissertation to defend the notion of commitment as important for the evolution of cooperation – this work has already been done, notably by Schelling (1960) and Frank (1988). Rather, I seek to show that the *coevolution* of new forms of commitment and new forms of cooperation was particularly important for our evolutionary trajectory toward our pervasive prosociality. In the following chapters, we will see how different forms of commitment work to secure mutually beneficial cooperation in different environments, and how the cooperation enabled by earlier forms of commitment created the selective environment for ever more effective forms of commitment to evolve. First, we must discuss the relationship between commitment as an explanation of cooperation and other existing theories.

1.6 Relationship of commitment to other theories of cooperation

The contribution in this thesis concerns the coevolution of commitment and cooperation. However, we should not confuse commitment with other models of cooperative behaviour. Prominent

explanations of the evolution of cooperation have included group selection, kin selection, reciprocal altruism, punishment or tit-for-tat models, indirect reciprocity, and secret handshakes, among others. Those which appeal to direct benefit, and are therefore explanations of cooperation in terms of mutual benefit, are reciprocal altruism, tit-for-tat, indirect reciprocity, and secret handshakes. Those which appeal to indirect fitness benefits, and are explanations of cooperation in terms of altruism, are kin selection theories. Group selection and cultural group selection generally include elements of both. Since my account is one of direct benefit, I will not address its relationship to explanations via indirect benefits as the difference should be clear. However, it is important to recognise that many explanations of the evolution of cooperation can be operating in tandem. That commitment secures cooperation in some contexts does not preclude that kin selection is also at work. More will be said on the relative strength of these theories in explaining cooperation in Chapter 2. In this section, I merely seek to show that commitment is a distinct mechanism and is not subsumed by other theories of direct benefit.

It is worth noting that many of the existing explanations of the evolution of cooperation have appealed to correlation devices – that is, mechanisms which increase the chance of cooperators interacting with other cooperators (Skyrms 1996). For example, kin selection provides a basis for altruists to interact with other altruists since they selectively interact with those who share their genes, either through familial bonds or due to limited dispersal. Secret handshakes (Robson 1990) and the greenbeard effect (Hamilton 1964; Dawkins 1976) are also a means of correlating interaction between cooperators via identifiable phenotypic markers. Reciprocal altruism and tit-for-tat strategies ensure cooperators interact with like kind because they withhold cooperative behaviour from non-cooperators. Indirect reciprocity facilitates cooperators interacting with like

kind since it involves discriminating partner choice enabled by information-sharing concerning the past record of interaction. Mechanisms by which cooperators identify and preferentially interact with other cooperators explain the stability of the evolution of cooperation where such a strategy is vulnerable to defection. This dissertation relies upon another means of correlating interaction: commitment. The theories which commitment most closely resembles are costly signalling, indirect reciprocity, secret handshakes and reciprocal altruism, so I will address its relation to these theories in turn, and show that it is distinct.

The costly signalling hypothesis is not a theory of the evolution of cooperation but proposes that some difficult-to-explain behaviours (such as bird ritual dances and early hominin hunting) are explained by agents signalling their phenotypic quality to reproductive partners, benefitting the sender of the signal (Hawkes 1991; Hawkes & Bliege Bird 2002; Bliege Bird et al. 2001). It therefore sounds similar to commitment in that commitment involves signals which benefit the sender. Costly signalling can be divided into (at least) two parts. First, there are hard-to-fake signals. These are signals that are not fakable in that they convey information about an underlying state or trait using a method that is unavailable in the absence of that underlying state (Brusse 2020). Signals of this sort might be made hard to fake by physical features of the agent or the cognitive load involved in making the signal appear credible (Frank 1988). However, I do not rely on commitment signals being unfakable. They are indeed fakable and there will be other means to ensure their credibility in the face of potential deception.

Second, there are signals which are costly-to-fake in the sense of involving a higher cost for the signaller who is of lower quality or is deceptive, in order to make these signals credible on the part

of honest agents.[28] While some commitment signals can be costly in this way, many are not. Consider, for example, uttering a promise in natural language. There is no cost asymmetry for honest and dishonest agents in issuing this signal which ensures that it is honest. Rather, the costs which ensure that the credibility of commitment signals can be maintained are largely *social,* imposed by another party, and are *contingent* upon defection. That is, they do not occur unless the agent reneges on the commitment. In Chapter 5, I argue it is more likely that legal and religious commitments involve upfront and intrinsic costs such that deceptive agents are deterred from making the commitment. This is one way in which there is a cost differential between high- and low-quality agents – high quality agents can afford to pay the investment cost. So a commitment may involve a costly signal or it may not – it is not the same as costly signalling.[29]

Indirect reciprocity explains the evolution of cooperation by way of reciprocation on the basis of one's reputation (Alexander 1987; Nowak & Sigmund 1998; Nowak & Sigmund 2005). The basic premise of indirect reciprocity is that an agent will not get their back scratched if it becomes known that she never scratches another's. Here, reciprocation may come from someone other than the recipient of the cooperative behaviour since reciprocation is based on one's record of past behaviours. The record of past behaviours comes from direct observation as well as reputation sharing. Gossip and reputation are indeed important in the commitment framework. Often, what makes a commitment change sender payoffs at the level of ultimate causation is its reputational

[28] In particular, costs and benefits must be distributed so that lies, but not honesty, are priced out of the evolutionary market (Brusse 2020). That is, the cost of signalling that one is a high-quality by a low-quality agent must be greater than the average benefit of being treated as a high-quality agent, while the average cost of the honest signal for a high-quality agent must be less than her ensuing benefit.

[29] Brusse (2020), however, argues that signals could also be costly to fake as a result of punishment. If we adopt this more expansive understanding, punishment (primarily social exclusion) renders all of commitment signals which I discuss costly-to-fake.

consequences. At the proximate level, fear of reputational damage may also motivate an agent to follow through on her commitment.

However, commitment involves more than just reputation. First, indirect reciprocity models do not involve pre-play signalling that changes a receiver's expectations. Credible pre-play signalling and following through on a commitment can alter an agent's reputation by affording her an opportunity to recover her tarnished reputation. As such, commitment goes beyond the indirect reciprocity account. Second, though reputation may be one of the ways in which commitments are enforced, there are many other means, too. Commitments can be internally enforced by way of emotions, or externally enforced either contractually or physically. Indeed, in Chapter 5, we see that commitments can aid cooperation without access to the reputation of the agent's potential partner at all. This will be so if the commitment can be made to be forcibly followed through by a third party.

Secret handshakes are pre-play signals which, if sent by both parties, result in cooperation (Robson 1990). This is the game-theoretical equivalent of the "greenbeard effect". The key idea is the introduction of mutations which involve possession of an observable characteristic which, by supposition, has zero inherent cost. This characteristic serves as a signal. Mutants are able to recognise the presence or absence of this signal in potential partners and condition their choice of strategy accordingly. The addition of a secret handshake mutant into the 2x2 Prisoner's Dilemma yields a 3x3 game in which the group-preferred cooperative outcome is the uniquely evolutionarily stable strategy. In being a pre-play signal which aids cooperation, commitment resembles a secret handshake.

However, commitments are not synonymous with pre-play signals. Pre-play signals will be commitments only when there is a *cost to reneging.* So the commitment framework offers an explanation of *why* we are motivated to follow through on our signalled intent rather than assuming the signal is characteristically held by individuals of a particular cooperative type. Further, in the secret handshake account, that the mutants possess this characteristic is accidental whereas, in the commitment account, it is evolved up. Finally, the commitment framework helps us make sense of how credibility might be maintained in the face of deceptive signals. In contrast, Robson (1990) states that those agents who use the secret handshake deceptively to exploit cooperators win out in the long run.[30]

Reciprocal altruism is a theory of the evolution of cooperation in which individuals pay a current cost for the benefit of a social partner's reciprocation (Trivers 1971). This reciprocation may take place immediately, as in food-sharing, or later in time, as in bird warning calls. It is represented strategically as the tit-for-tat strategy in game theory. Although the theory is adequate to explain many of the cooperative behaviours mentioned in this dissertation, it misses some important details which commitment captures. Not only this, but there are important conceptual differences.

In particular, reciprocal altruism does not involve pre-play signalling which alters receiver's expectations so cannot be used to secure cooperation in a simultaneous move game since, if there is no pre-play signal, the receiver has no information on which to base their expectation of the

[30] Subsequent work has shown that such invasion is not always detrimental to the persistence of effective secret handshakes (Wiseman & Yilankaya 1999; Grégoire & Robson 2003; Santos et al. 2011).

sender's cooperation. Reciprocal altruism is therefore applicable to games in which there are two temporal stages, as in a sequential game, or where there are repeated games. Yet commitment can serve to secure cooperation in simultaneous-move one-shot games (consider legal commitments which do not depend on reputational enforcement mechanisms). Commitment also differs from this theory in the proximate mechanisms, explicated in Trivers' initial work. In particular, Trivers (1971) points to the emotional rewards we experience for friendships, feelings of gratitude, sympathy, guilt and moral aggression. Note that some of these may also play a role in commitment, but importantly, proximate mechanisms related to hypocrisy and ostracism hold a central place.

A final note: one might wonder how the operationalisation of commitment differs from a combination of pre-play signalling accompanied by punishment. The two are logically distinct but not mutually exclusive. First, a commitment differs from a pre-play signal in that it necessarily involves a cost to reneging which motivates the sender to act in line with their signal. Second, note that, at some level of description any explanation of the evolution of cooperation that rests on correlated interaction can be understood as operating by punishment of defectors (at least in the sense of excluding them). Commitment differs in the mechanism by which this is achieved – it is via signalling her intended cooperation that the sender opens herself up to potential exclusion if she subsequently defects. Also note that commitment is useful in a pure coordination game such as the Stag Hunt, as well as public goods games with the potential for free-riding. That is, even when there is no "defector" as such in the Stag Hunt, commitment allows us to settle on the more risky but more profitable option of hunting Stag. This will not be true of other accounts of punishment-based correlated interaction, which rely on the Prisoner's Dilemma set up.

1.7 Outline of the dissertation

The purpose of this dissertation is to draw our attention to an overlooked and significant part of the evolution of human prosociality – the coevolution of commitment and cooperation. In the remaining chapters, I show how different methods of undertaking commitments in our evolutionary history have enabled more sophisticated forms of cooperation over time which, in turn, create the selective environment for the evolution of increasingly effective commitments. I detail the emergence of and consequences of four types of commitment: pre-linguistic commitment via participation in shared activity, explicit and implicit linguistic commitment, "moralised" commitment, and institutionalised commitment.

In Chapter 2, I argue that participation in group hunting is an instance of commitment. It is a signal of intended cooperation which changes the sender's relative payoffs for following through and changes the receiver's expectations of her doing so. The commitment raises the cost of defection by making the reneger vulnerable to potential exclusion from the benefits of current and future cooperation, rendering following through on one's commitment the fitness-enhancing option. Not only this, but it is a commitment which may itself generate future opportunities for beneficial interaction via the creation of a social bond. The bonding experience, associated emotions or dispositions to punish plausibly act as the proximate psychological mechanisms by which sender defection becomes undesirable at an agent level, motivating the agent to act in line with her signalled intent. Further, the expectations of the receiver change in virtue of this signal of participation and intended cooperation, allowing receivers to identify trustworthy partners and thereby acting as a means by which the receiver can secure correlated interaction with other

cooperators. Ultimately, this allows cooperative tendencies to spread in the population with reduced risk of invasion from defectors.

I also argue that successful hunting, and other factors such as cooperative breeding, enabled new forms of cooperation: the development of more sophisticated tools; the division of labour; specialisation; new cognitive capacities for cooperative interaction in infants; an expansion in terrestrial habitats; better territory defence, and information sharing. These innovations, along with other ecological factors, led to the formation of larger, multi-level groups and this selective environment favoured the development of linguistic communication to coordinate among new group members. Not only this, but the cognitive preconditions to the emergence of language – for example, the ability to make perspectival representations, socially-recursive inferences and to self-monitor – would themselves have been selected for by shared activities such as hunting. In this way, our previous practices of commitment contributed to a cooperative environment in which linguistic information sharing was both exceptionally useful and they set the foundation for such communication to evolve from earlier proto-languages.

In Chapter 3, I discuss how the advent of language enabled correlated interaction based on reputation and a new form of commitment – linguistic commitment. Linguistic commitment can be either explicit, as in promising to aid another, or implicit on the basis of gossip about a third party. Gossip about agents in one's social network allows cooperators to identify potential partners for mutually beneficial interaction via access to their reputation. I argue that gossip concerning a third party's adherence (or lack thereof) to a norm can be taken as an implicit commitment. It constitutes a commitment to behave in a similar/dissimilar manner in the context of the receiver's

beliefs about the agent's cultural background and role. That is, such gossip statements will operate as commitments as long as the receiver believes the sender to share the same cultural norms and social role as the agent about whom they are gossiping. If this is so, she will view her interlocutor as a less desirable partner should she act otherwise than she has signalled via her gossip, meaning there is a cost to acting against one's signal, characteristic of a commitment. In this way, normatively-laden gossip acts as an implicit commitment which, again, both changes the relative payoffs of the sender and expectations of the receiver.

As with commitment via shared activity, both types of linguistic commitment introduce a fitness cost to reneging for the sender and facilitate the identification of cooperative partners for the receiver. Here, the commitment does not itself generate opportunities for future beneficial interaction since it does not (necessarily) involve a shared bonding experience. Rather, opportunities for beneficial interaction – with the same agent and with others – are available in virtue of an agent's reputation. In this context, the ability to commit and follow through on one's commitment may act as a means by which one's reputation can be altered since news of one's actions will spread. The proximate mechanisms underpinning an agent's motivations to cooperate include those previously outlined, but now also extend to fear of ostracism and concern for one's reputation in a social network, since defection from a commitment is subject to gossip.

I suggest that these new linguistic forms of commitment facilitate cooperation more effectively than commitment via shared activity. In particular, language allows us to extend cooperative interaction to new partners with whom we might not have had opportunities for shared activity but whose reputation is accessible to us. This allows us to accrue further fitness benefits from

interaction with more cooperative agents. Language enables more precise communication, resulting in more grounded expectations of behaviour. It also allows us to make conditional promises which may increase the credibility of a commitment. Furthermore, via linguistic commitment, we expand the range of matters on which we can undertake commitments since language is flexible enough to matter for cooperation in new contexts and can be used to communicate commitments concerning issues which are spatially and temporally remote from the current environment. This also provides more opportunities for assessment of a partner's reliability since it provides more scenarios in which they can demonstrate themselves as trustworthy or untrustworthy, thereby aiding in correlated interaction among cooperators. As the realm of our cooperative enterprises extend, selection favours an additional means of discriminating among potential partners and making oneself more attractive to others.

In Chapter 4, I discuss how, at some point in our evolutionary history, we began to consider some norms as universally applicable or externalised.[31] To directly assert one's attitude toward an externalised norm or to gossip about another's violation of an externalised norm is to suggest one's own commitment to that norm, since it is considered universally applicable and therefore relevant to the sender's behaviour. I call these commitments "moralised commitments" for ease, though the category of the externalised will not always perfectly apply to all and only moral norms. The evolution of this capacity to externalise norms has deep ties to previous commitment practices. I argue the empathy typically characteristic of moral cognition requires taking on another's perspective, and such capacities likely evolved in order to facilitate action in joint tasks such as hunting. The affective mechanism which motivates helping behaviour is also built upon the ways

[31] See Stanford (2018a).

in which early cooperative breeding has perceptually tuned us to emotions. Furthermore, externalisation of norms became an important mechanism for partner choice in the context of increasing group size and complexity, which was partially a result of this earlier small-scale collaboration. In this way, our commitment practices again coevolved with cooperation. Earlier forms of commitment and cooperation both provided the cognitive prerequisites, and contributed to the selection pressure, for the development of externalised norms. This, in turn, allowed for a new form of commitment to evolve to facilitate cooperation in an increasingly complex world.

With moralised commitments, further proximate psychological mechanisms become relevant to an agent's motivation to cooperate, for example, the phenomenological character of universal obligations, our attitudes toward moral norms and transgressions, and our dispositions to view hypocrisy with outrage. Moralised commitments offer a number of fitness advantages over and above the linguistic commitments of Chapter 3. First, externalised norms make implicit commitment independent of the cultural background and role of the sender of the commitment signal, meaning these commitments are much easier to make and to identify. Second, by making a moralised commitment, the sender not only advertises her trustworthiness but simultaneously reveals to the receiver that she is willing to exclude others for not acting in like manner. She thereby secures better correlated interaction. Finally, there is psychological evidence that points to the strength of moralised language and gossip in changing receiver expectations of a sender's future action.

In Chapter 5, I go on to argue that the cooperative environment made possible by commitment via shared activity, explicit, and implicit commitment set the stage for a fourth form of commitment

to evolve: institutionalised commitment. Explicit or implicit commitments made in a public setting would have afforded onlookers the opportunity and incentive to enforce these commitments in order to bolster their own reputation within their subgroup. Herein, I suggest, lies the evolution of third-party punishment with respect to commitment. The transition to more hierarchical societies after the Neolithic revolution would have added additional resources to third-party punishment through elite enforcement. At the same time, institutions were becoming increasingly formalised (for example, clearly codified and premised on organisational goals) in response to the complexity of cooperative life. This allowed third-party punishment to become organised, lower cost, and common knowledge. With the modern policing bodies of governments and courts, the threat of third-party punishment is a clear constitutive part of institutionalised commitment and this had great effects for the extent of human cooperation.

The Neolithic revolution and transition to agricultural and hierarchically organised life also has important ties to our previous forms of commitment. In particular, small-scale shared activities such as hunting contributed to growth of the group size and a division of labour. Linguistic commitment allowed us to coordinate on complex collective action problems which demand plasticity of response. Yet with the advent of settled, agricultural life, interaction with strangers was common and reputational means of enforcing commitments became stressed. This created selective pressure for the strengthening of third-party punishment to act as a means of enforcing commitments. With institutionalised commitment, the state may impose a fine, imprisonment or reparations equivalent to fulfilment of the commitment. Institutionalised forms of commitment widen the scope of one's potential partners at the same time as lowering the costs of enforcement relative to the punishment enacted. In some cases, this leads to a lowering of the cognitive demands

of trust. This is because agents may not need extensive information about the reputation of an agent in order to consider it worthwhile to interact with new partners. As long as amends can be forcibly made, an agent's prosociality is institutionally protected.

The trajectory of this relationship is illustrated in Figure 5. The arrows in this diagram represent both a proposed chronological order and causal relationships. At times, the causal relationship will be robust, at others, the previous stage is a necessary condition for the emergence of a later stage but does not itself select for it. More will be said on this distinction in the dissertation.

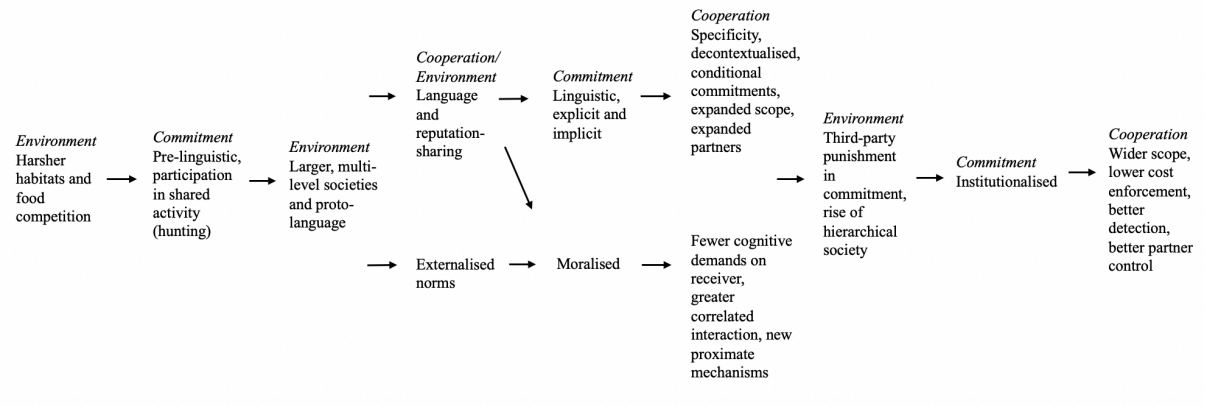

Figure 5: The coevolution of commitment and cooperation.

This dissertation argues that the evolution of modern human prosociality was in large part due to the coevolution of commitment and cooperation over our evolutionary history. Commitments have allowed us to correlate interaction effectively by simultaneously changing the incentives of the sender of the commitment signal and the expectations of the receiver. At each stage in our evolutionary history, new forms of commitment have evolved to better facilitate cooperation in a world where our interactions have expanded in scope or fashion. This is not to say that the

commitment account renders implausible other accounts of the evolution of cooperation. It is consistent with all that I have said that kin selection, reciprocal altruism or correlated interaction by other means have played some role in the evolution of human cooperation. My point is rather that we miss many important details of how we actually came to be so cooperative when we focus only on these theories.

I hold the commitment framework precisely characterises behaviour at the right level of generality in order to unify seemingly disparate and difficult to explain behaviours – promising, moral gossiping, religious oaths, among others. All involve pre-play signalling which changes receiver expectations and sender motivations. Describing a model that fits behaviour at the right level of explanation is a worthwhile endeavour. As an analogy, one could use "hot" and "cold" to describe temperatures but oftentimes this is too coarse-grained an explanation to be useful. Alternatively, one may use an explanation that is too detailed and is thus cumbersome, for example, referring to the kinetic energy of each molecule of a substance. The theory of commitment is intended to explicate behaviour at the most *useful* level of description, analogous to providing the Celsius or Fahrenheit scale for temperature. This explanation is useful because we see that it is in the *details of the commitment mechanism* that new forms of securing cooperation are more advantageous than previous forms.

To elucidate, a linguistic commitment is more effective than commitment via shared activity since language more effectively changes receiver expectations due to its specificity, its decontextualised nature, and due to the ability to strengthen promises through specifying collateral. Moralised commitments will also be fitness-enhancing relative to making non-moralised linguistic

commitments since sender motivations are now strengthened by the phenomenology of externalised norms at the proximate level, and receiver expectations of both the sender's actions and sender's expectations of others are clearer. Institutional commitments offer fitness advantages over and above the previous forms of commitment by increasing the cost of defection through third-party punishment, which can more effectively change sender motivations. They may also serve to make receiver expectations more precise due to the (often) codified nature of these commitments.

The thesis that the evolution of modern human prosociality is largely attributable to the coevolution of commitment and cooperation is also novel. One might worry that, much like the other theories of the evolution of cooperation, it relies upon cooperation ultimately providing some benefit to the cooperator, or that it relies upon the exclusion of defectors, which seem to be features of theories such as reciprocal altruism, punishment, indirect reciprocity and secret handshakes. However, notice that at some level of description *any* explanation of the evolution of cooperation that purports to explain mutual benefit, and not genuine altruism that is costly to the agent, will appeal to direct benefits to the cooperator, so this should not be a mark against the originality of the thesis. Similarly, any explanation that relies on correlated interaction will involve the exclusion of defectors, since this is all that correlated interaction amounts to. The merit of presenting this new hypothesis is to explicate *by what mechanisms* the cooperator benefits and the defectors are excluded.

My account of the evolution of cooperation via commitment also differs from previous accounts on the role of commitment in cooperation (Schelling 1960; Frank 1988; Hirshleifer 1984; Gilbert

2013; Hans 2013; Back 2007). Rather than showing a proof of possibility for the role of commitment in cooperation, I have detailed a how-plausibly story, grounded in empirical evidence, of how *new* forms of commitment have evolved *over time*. Not only this, but I have argued that there is a *coevolutionary relationship* between our forms of commitment and our forms of cooperation. That is, I showed how different methods of undertaking commitments in our evolutionary history have enabled more sophisticated forms of cooperation over time which, in turn, create the selective environment for the evolution of increasingly effective commitments. None of the aforementioned accounts of commitment deal with the evolution of new commitments over time, or the coevolutionary relationship between commitment and cooperation.

Finally, alongside highlighting an important feature in the evolution of modern human prosociality – the coevolution of commitment and cooperation – this dissertation also elucidates how and why we commit in the various ways that we do. It offers an explanation for how we are able to commit in the absence of language, why we believe others' commitments, how gossip can be informative for others' future actions, the effect morality has on our ability to make promises to one another, and how we have strengthened our means of commitment through institutionalisation. Each of these developments is a response to a changing cooperative environment. We have continued to find more effective ways to ensure that we make choices that serve our long-run interests and which signal our trustworthiness to others.

Chapter 2: Commitment via shared activity

We begin our account of the evolution of cooperation with a pre-linguistic form of commitment based on shared activity. I argue this form of commitment was involved in the cooperative hunting of early hominins. Not only this, but hunting, along with other factors, set the stage for more complicated forms of cooperation to evolve. Commitment facilitates cooperation via two effects. First, it alters the sender's relative payoffs such that cooperation achieves better fitness consequences in the face of potential exclusion from the benefits of cooperation and from beneficial future interaction. Second, it transmits information about the intended actions of the sender to the receiver, allowing the receiver to identify and choose trustworthy partners. Opportunities for future interaction are themselves generated by participation in shared activity. Here, honest commitments are stable as a result of direct monitoring and the social costs of exclusion, allowing cooperative tendencies to spread in the population with reduced risk of invasion by defectors.

In this chapter, I have two aims. The first aim is to argue that commitment provides a good model of early hominin group hunting, and to support this explanation with empirical evidence. There are likely many other shared activities in our ancestral environment that possess the same formal features of commitment, for example, in gathering or transportation of materials. I focus on hunting in this dissertation only as an illustrative and important case. The second aim is to show how this form of cooperation contributed to the formation of larger, multi-level societies, constituting the first stage in our coevolutionary story. The outline of this chapter is illustrated in Figure 6 below.

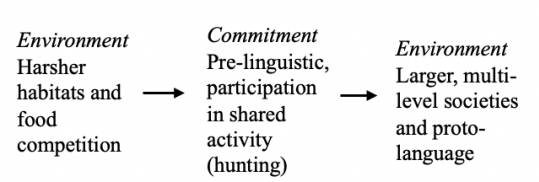

Figure 6: The coevolution of commitment and cooperation.

2.1 The emergence of group hunting

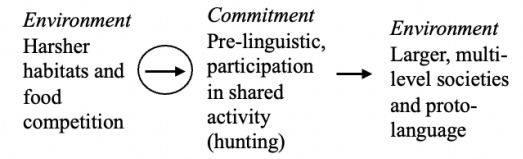

Figure 7: The emergence of pre-linguistic commitment via shared activity.

After the emergence of the genus *Homo* approximately two million years ago, global cooling and drying trends created an expansion of open environments. The change in habitat was coupled with an increase in terrestrial monkeys such as baboons who outcompeted early hominins for their usual means of subsistence in fruits and vegetation, creating pressure for early hominins to find a new foraging niche (Tomasello et al. 2012). Choosing to forage individually was no longer optimal. Hunting "stag" became more important than hunting "hare" and stag hunting required coordination, especially since long-range weaponry did not exist until the late Middle Stone Ages. Archeological evidence suggests that hominins engaged in ambush hunting as early as 2-1.8 mya (million years ago) (Pickering 2013; Bunn & Pickering 2010). There is evidence of food being brought back to a home base from at least 400-200 kya (thousand years ago) (Stiner et al. 2009).

Tomasello and colleagues (2012) hypothesise that the transition between non-cooperative hunting and food sharing to cooperative hunting and food sharing may have occurred first because of the need for scavenging of carcasses killed by other animals. He suggests individuals who were tolerant of conspecifics cofeeding on the same carcass would do better, for example, by not paying the cost of driving others away or risking losing the carcass in a dominance battle.[32] Thus, a key factor in the selection for cooperative tendencies was increasing interdependence. Indeed, experiments have shown that chimpanzees who exhibited tolerance in food tests were more likely to solve novel cooperative tasks (pulling a tray together for a food reward) (Melis et al. 2006). It is also instructive that bonobos, who are naturally more tolerant of food sharing, perform better in cooperative food retrieval tasks than chimpanzees (Hare et al. 2007).

Another key factor in the emergence of hunting as a collaborative activity was the erosion of the traditional primate dominance hierarchy, hypothesised to be occurring in the Pliocene or early Pleistocene (Sterelny 2021a). It is possible earlier forms of collective defence may lay the cognitive and social groundwork for such changes. In chimp societies, high value foods are stolen by dominant individuals. This disincentivises investment in collaborative activity. However, by the Pliocene, there was evidence of reduced male-male competition in the hominin lineage. Sterelny (2021a) argues that the suppression of the dominance hierarchy in the hominin lineage was partly due to the rise of weaponisation, and the evolution of greater social intelligence and impulse control.

[32] See also Sterelny and Planer (2021a) and Sterelny (2012) on how scavenging was an intermediate step between the feeding habits of the australopithecines and early hominins.

Chimpanzees lack the cognitive and communicative tools to permanently usurp an alpha. The rise of weaponisation in the early Pleistocene changes this, as hominin resentment storms would likely be much more dangerous, further disincentivising monopolisation and providing the foundation for cooperation to get off the ground (Bingham 1999, 2000). Moreover, once any kind of collective activity emerges, this activity would select for coordination and impulse control skills which provide the groundwork for a transformation of simple reactive attitudes toward dominance into sustained coalitions with enduring motivations (Sterelny 2021a). This feeds back into successful cooperation by further suppressing dominance hierarchies. So we have an explanation of the key factors that led to the emergence of collaborative hunting: ecological changes, scavenging selecting for tolerance, suppression of the dominance hierarchy, weaponisation and impulse control skills.

Despite its likely initial emergence through scavenging, we will see that there is still a free-rider problem in cooperative group hunting that is in need of explanation. When agents engage in a hunt, it is in the interests of each to free-ride off the contributions of others. This will be true as long as there is some cost to participation and there is a threshold level of contribution which makes a hunt successful, beyond which arises the possibility of free-riding. In other words, when the hunt is well modelled by a Threshold Public Goods game or an *n*-player Stag Hunt. If the cost of participation is low and the probability of any particular agent being the difference-maker in the success of the hunt is high, the agent is incentivised to cooperate. However, often, these conditions will not hold. The hunting party will not know in advance how many individuals are needed to secure a particular prey since the prey's response to the particularities of its surroundings might be different. Prey will also differ in age, health, or size. As a result, hunts may end up with more people than is

strictly necessary to catch a particular prey, meaning that opportunities for free-riding exist. Not only this, but if hunting involves many different types of actions and phases (tracking, surrounding, attacking), it is likely that not all of the hunting party are needed for each of these actions. It is possible that an agent contributes to surrounding the prey but free-rides when the time comes to attack it – all agents may contribute in some way, but the question is how much they do at each point.

In the archeological literature, we see that in the early Acheulean (c. 1.75-1.6 mya), hominins engaged in systematic and regular exploitation of megafauna, which were both faster and stronger than the scavenging and hunting targets of the Oldowan period (c. 2.6-1.75 mya) (Domínguez-Rodrigo & Pickering 2003). This shift represented an important change in hominin predatory capabilities where, earlier, hominins might have focused on smaller quarry that could be killed using hands, teeth, or perishable weapons (Domínguez-Rodrigo & Pickering 2003). If this change required the introduction of more members in a hunting party in order for a hunt to be successful, we may eventually see the introduction of redundancy in numbers, particularly as long-range weaponry develops or there is better transfer of skills to future generations.[33] Free-riding of this sort is most likely to be seen in ambush hunting. In ambush hunting, as opposed to endurance or encounter hunting, agents are spatially remote from one another so may not have direct visibility over each other's actions. Yet, free-riding behaviour could not have been prominent among early hominins or cooperative hunting would not have persisted. When individual defection may be optimal in this context, how is cooperation sustained?

[33] This shift in predation practices also coincided with growth of the group size, suggesting that there might have been a coevolutionary relationship between pressure to hunt big game in larger parties and the development of larger groups. That is, larger groups require more resources and these resources in turn lead to the development of larger neocortexes which support increased group cohesion among more members, allowing successful group hunting (Dunbar 1998).

2.2 Costly signalling

Hawkes (1991) famously argued that the reason men partake in a collaborative hunt is to signal their phenotypic quality to potential reproductive partners. She holds collaborative hunting is not a form of family provisioning, but a means of "showing off". This suggestion was later connected to game theory as an instance of costly signalling (Hawkes & Bliege Bird 2002; Bliege Bird et al. 2001). The costly signalling literature tells us that agents have an interest in advertising their quality if it is not directly observable, either for sexual access or resource attainment. Such signals may take the form of risky displays or energy-intensive activities. In this context, it would also be profitable for low quality individuals to deceive others about their quality. However, in general, the receiver of the signal is able to determine when senders are honest about their quality because of the cost entailed in signalling – a low quality agent could not afford to signal (Zahavi & Zahavi 1997; Grafen 1990). Some examples of costly signalling in human behaviour are feasts hosted by village leaders, gift-giving, and displays of courage in intercommunal conflict.

Under the costly signalling view, the primary returns to hunting are social status and sexual access rather than provisioning the hunter's family. However, Sterelny (2012) argues that the costly signalling view is mistaken in most hunting situations.[34] This is because hunting is a collective activity in which all members are signalling; it is difficult to observe individual contributions; and in which costs do not fall differentially on the less-skilled. The whole group may do less well with lower quality hunters but it is not clear the individual hunter does less well if he is low-quality,

[34] With exceptions, such as in the case of Meriam turtle hunters (Smith et al. 2003).

particularly if individual risk is mitigated by the use of long-range weaponry. Furthermore, we would expect individuals to withdraw from costly hunting if they were less skilled – it would only advertise their low quality and they would pay the costs of signalling – but we do not see this in reality.[35] We do not observe within-group differences in male attitudes to hunting which match differences in male quality (Sterelny 2012). Furthermore, Sterelny argues that hunting is a good form of provisioning, lending support to the idea that its role is not primarily for signalling hunter quality. In a study of ten forager societies, it was found that, on average, male hunting generates 68 per cent of the group's calories, compared to female foraging which generates 32 per cent, as well as providing 88 per cent of total protein sources compared to just 12 per cent from gathering (Kaplan et al. 2000).

Gurven and Hill (2009) emphasise the empirical arguments against the purely costly signalling view. They focus on the fact that proponents of the costly signalling view hold that familial provisioning cannot be the motivation for hunting, else male foraging patterns would more closely mimic those of women. They note that this view depends on four key assumptions: (1) that men forage for large-package-size items even when alternative foraging strategies yield a higher long-term average food value (in order to signal quality to potential mates rather than to provision); (2) high-variance daily acquisition activities (in hunting) cannot effectively provision offspring; (3) food transfers by hunters are not paid back later in currencies that directly affect familial welfare,

[35] There is a mechanism by which all agents, even those that are low-quality, signal. This is known as the "full disclosure principle" (Frank 1988). To understand how this works, let us borrow an example from Frank (1988) on frog croaking. Female frogs prefer to mate with the male frogs with the deepest croaks, but these can only be achieved by large-bodied frogs. Once the deepest croak is heard, we must ask: why do the smaller frogs continue to croak, as surely this would advertise their lesser quality? The answer is that the second-largest frog croaking advertises that he is larger than the average. If he does not croak, the female may assume he is smaller than he is. The third largest frog should croak for just the same reason. So all agents signal honestly and signalling does indeed reflect quality. However, Sterelny (2012) argues that the full disclosure principle neglects costs – there will be a point where the costs of croaking outweigh the benefit to be gained by advertising.

and; (4) women prefer gathering over hunting only because of its higher reliability and productivity. All of these assumptions are meant to underly the argument that hunting is not motivated by familial provisioning, but rather signalling one's quality.

It was previously thought that studies from the Ache, Hadza and Hiwi supported (1) and (2), which state that other foraging strategies would be more productive for men in the provisioning of their offspring. In particular, Hawkes (1991) argues that Ache men could gain energy at higher rates by extracting palm products (Hawkes 1991). Gurven and Hill (2009) find that this is mistaken due to laboratory error in assessing edible portions of palm fibers and recent experimental evidence shows that only 1/35 encounters with palms involve exploitable starch. They argue mixed hunting and collecting patterns would produce comparable calories per unit time as would exclusively foraging for palm fiber. If this is the case, hunting could be a good means of provisioning, rather than of signalling quality. Hawkes and colleagues also argued that Hadza large game hunters would be more productive should they hunt small game, suggesting that the reason they opt for high-variance acquisition activities such as hunting large game is to signal male quality (Hawkes et al. 1991, 2001). However, recent data has found Hadza large game hunting produces nearly twice the energetic-gain rate as gathering and up to four times more than small game (Gurven & Hill 2009).

In response to (3), Gurven and Hill (2009) provide evidence that meat transfers are reciprocal. For example, in the Hiwi, the amount of meat shared was the strongest predictor of the amount of meat returned by the same family over a four-month sample period (Gurven et al. 2000), this was also true of the Pilaga (Gurven 2004), Yanomo (Hames 2000), Dolgan and Nganasan (Ziker & Schnegg 2005). Sharing patterns in women of the Ache and Hiwi are also similar to those of men, and this

is paid back in a form useful for parental investment (reciprocal food-sharing) rather than mate access, undercutting the thesis that costly signalling is the motivation for male hunting. The costly signalling hypothesis also suggests that women avoid hunting since the payoffs are low or unpredictable, as in (4), but women in the Agta also hunt, and Ache and Hiwi women participate in male hunting activities in supporting roles (Gurven & Hill 2009). As such, the costly signalling hypothesis is not well supported by ethnographic evidence. Provisioning offers an explanation for why men hunt, but we still face a free-rider problem. It does not necessarily require the effort of an individual hunter to ensure his family is provisioned as long as the hunting party meet a threshold level of engagement for a successful hunt and the individual hunter's lack of contribution goes unnoticed.

2.3 Committing to hunting

So what explains the stability of group hunting even in the face of this free-rider incentive? Recall from Chapter 1 that the evolutionary explanation of cooperation I am putting forth is one of mutual benefit. That is, my account explains the stability of cooperation by way of direct benefits to the cooperator. I argue that cooperation becomes directly advantageous to an agent in virtue of commitment. A commitment is a signal of intended cooperation which, if carried through, confers future benefits to the sender by way of increasing access to further beneficial interaction. At the same time, it allows receivers of the commitment signal to identify their potential partner as trustworthy or untrustworthy. In what follows, I present the account of Sterelny (2012) on how participation in group hunting is a commitment and offer a critique. I then present how participation operates as a commitment under my operationalisation of commitment.

Sterelny (2012) argues that participation in collective hunting is a *commitment* that itself generates future opportunities for beneficial interaction by way of creating a valuable relationship. To see how this works, note that because individuals "have invested time in relationships, these relationships change payoffs in triggering situations – the cost of defection has been driven up since defection risks fracturing trust and forfeiting the profit from the investment" (Sterelny 2012: 117). Opportunities for future cooperation in virtue of this social bond increase the fitness cost of defection from collaborative activity, since defection risks punishment by exclusion. The profits that the agent risks forfeiting are both the fruits of mutually beneficial cooperation in future interactions and opportunities for future interaction – if one lags behind in the hunt, one will not receive a share of the acquired resource and one will be a less desirable partner for further cooperative interaction. Furthermore, if profits are accumulated through multiple different cooperative interactions – for example, in foraging, hunting and collective care of offspring – the cost of defection is further increased, as is the corresponding fitness benefit of acting in line with one's signalled intent.

Sterelny argues the joint success of hunting results in an effect known as "arousal" which is particularly strong in demanding and emotionally-amplified shared experiences, and it is this which is the proximate mechanism behind the formation of social bonds and underlies the agent's motivation to act in line with her signalled intent. Note that the ultimate cause of cooperative behaviour via commitment is elucidated in terms of forgoing the fitness benefit from continued interaction with other agents – it is social in nature. The proximate cause of commitment behaviour, however, is a result of a change in personal utility due to arousal.

On Sterelny's view, commitment is a form of *niche alteration*. That is, commitment changes our environment such that carrying through on the commitment becomes the optimal choice. For example, tribal tattoos are interventions into our environment that make cooperation within the group the optimal option since they limit one's access to other social networks. Commitments are also *signals* since the niche alteration is public. In particular, then, hunting is an intervention which alters an agent's environment by changing the opportunities available to him. On this view, group hunting does not signal underlying quality but rather *creates* the conditions for trust incrementally as a result of costly, high arousal activity. Joint action in emotionally amplified shared experiences, such as combat environments, reinforces mutual bonds more so than joint action in calm environments (Sterelny 2012).[36] These activities then generate future opportunities for beneficial interaction by way of building trust, though the agents will have been unsure of each other's cooperativeness to begin with. Specifically, if an agent sees that another has cooperated with her in the past, she is able to engage in future beneficial activities with that agent with decreased fear of defection. All that is required is that there are future opportunities for beneficial interaction, and defection from the current cooperative activity risks punishment by exclusion. This means the cost of defection is driven up. So hunting is not primarily a signal of phenotype quality, but rather a signal of intended cooperation which constrains future choice by changing payoffs. Commitment signals incentivise hunting by raising the cost of defection from hunting activity, as to do so risks fracturing a profitable relationship.

[36] In order to explain the internal motivational structure of joint activity, Sterelny appeals to the work of Tomasello and others on shared intentions (Tomasello et al. 2005; Tomasello & Carpenter 2007; Tomasello 2008, 2009). He argues that joint attention is not only a description of several uniquely human cognitive capacities for cooperation, but also a *motivational state*. Experiments show that children (but not chimps) engage in joint activities even in the absence of an external reward, suggesting that the motivation to cooperate in joint activities is deeply ingrained in humans (Warneken & Tomasello 2009).

Sterelny notes that the external enforcement mechanisms presented in Schelling's (1960) original work would not have been available in our ancestral environment. That is, our ancestors did not have contracts, formal institutions, or public rituals to raise the cost of defection. However, Sterelny also rejects the operation of commitment by fully internal enforcement mechanisms such as emotions in Frank's (1988) account. He argues that "commitment mechanisms primarily depend not on subjective rewards and signals but on constructing environments that channel choice" (Sterelny 2012: 108). Despite the fact that Sterelny seeks to delineate his account of commitment from that of Schelling's and Frank's, I argue that they are not appreciably different. In order to understand why, we must again invoke the distinction between proximate and ultimate causation, which is not made explicit in Sterelny's work.

From the perspective of proximate causation, what serves to secure commitment in Sterelny's case is "arousal", which is an affiliative emotion that creates mutual bonds. This is analogous to the effect of emotions in Frank's account, which add internal utility or disutility to current options. Here, the current option is to pull one's weight in the hunt, rather than lag behind, and arousal is the proximate mechanism which incentivises this. Thus, commitment does rely upon "subjective rewards" at the proximate level. From the perspective of ultimate causation, what makes commitment fitness-enhancing is the possibility for future interaction – to commit and to follow through on this commitment allows access to future benefits in virtue of one's trustworthiness. Yet Schelling suggests a similar mechanism at play in the case of promises: "what makes many agreements enforceable is only the recognition of future opportunities for agreement that will be eliminated if mutual trust is not created and maintained, and whose value outweighs the

momentary gain from cheating in the present instance. Each party must be confident that the other will not jeopardize future opportunities by destroying trust at the outset" (Schelling 1960: 45). Thus, what secures promises from the perspective of ultimate causation for Schelling is not simply an external enforcement mechanism – it is the fitness consequences of acting in line with one's signalled intent. So, commitment on both accounts rely on creating "environments that channel choice" at the ultimate level. Thus, Sterelny's account accords with Frank's on the question of proximate causation of cooperative behaviour and with Schelling's on the question of ultimate causation.

There is one aspect of Sterelny's (2012) account of commitment via shared activity which *does* differ from the traditional literature, and I argue this divergence is unnecessary and obscures a useful distinction. Sterelny (2012) argues that some commitment signals change the motivations of both sender and receiver and, as such, they are not signals of the sort seen in classical game theory, where what changes for the receiver are her expectations rather than her payoffs: they are instead "Krebs-Dawkins signals" (Dawkins & Krebs 1978a, 1978b). They are *motivation-altering* rather than *information-carrying.* Dawkins and Krebs push back against the classical ethological view that cooperation is facilitated if signals are "efficient, unambiguous and informative" and where selection favours those actors who make it easy for reactors to read their internal state (Dawkins & Krebs 1978a: 289) Under their proposed alternative, the signaller communicates not in order to tell the receiver what the receiver wants to know, but to induce the receiver to do something that will benefit the signaller. They argue that "if information is shared at all it is likely to be false information, but it is probably better to abandon the concept of information altogether. Natural selection favours individuals who successfully manipulate the behaviour of other

individuals, whether or not this is to the advantage of the manipulated individuals" (Dawkins & Krebs 1978a: 289).[37]

I do not believe we need to reject the classical information-carrying account of signals in order to understand participation in group hunting as a commitment. Instead, we must recognise there are two parts to a commitment: one operates via altering the behaviour of the sender of a commitment signal and one operates via altering the behaviour of the receiver. Signals are indeed motivation-altering for the sender, but they are information-carrying for the receiver and the information concerns the intended actions of the sender. This is what allows her to identify the sender as a cooperator and engage in effective partner choice.[38] Sterelny's motivation for appealing to Krebs-Dawkins signals is stemming from his rejection of the costly signalling framework, where signals provide useful information about phenotype quality. However, he rejects too much. The commitment signal is still informative, but it is informative about what an agent will do in a strategic game, not about the agent's quality. The classical account which Dawkins and Krebs push back against – where cooperation is facilitated if information is shared and where selection favours those actors who make it easy for reactors to read their internal state – is exactly the view that Sterelny is espousing in his story of the emergence of trust. The actors in Sterelny's account are not being manipulated; the signals are intended to secure mutual benefit typical of informative signals.

[37] Examples include the colour patterns of Batesian mimics, the mimetic gape of the cuckoo chick, and the bluffing threat displays of moulting (and therefore vulnerable) stomatopods. One might wonder why the receiver would be selected to respond to the signal at all. To this, Dawkins and Krebs appeal to "mind-reading" – the potential to predict behaviour from the signal, which would be useful to the receiver (Dawkins & Krebs 1978b). This is similar to Barrett's and Skyrms' (2017) "cue reading". However, to appeal to this reveals that the distinction between manipulative signals and information transfer was not a useful distinction to begin with, since the receiver still makes use of the information in the signal regardless of whether it was the intended use of the signal.

[38] See also the evolutionary model presented in Chapter 1.

In fact, it seems we can only make sense of payoffs changing for the receiver in the way Sterelny suggests when we understand the receiver as herself a participant in the hunt – and therefore as a *sender*. This is because the payoff change is a result of potential exclusion from the spoils of the hunt and from future opportunities for interaction, and such opportunities arise in virtue of being a participant in the hunt and creating a social bond – in virtue of being a sender of the commitment signal. A receiver properly understood undergoes no increased cost of defection. In other words, on Sterelny's picture, there is no clear delineation between sender and receiver: all participants in the hunt are both and this is what pushes him to invoke the concept of motivation-altering Krebs-Dawkins signals instead of the information-carrying signals of classical game theory.

By making the delineation between sender and receiver clearer and emphasising the distinction between the motivation-altering and information-carrying aspects of a commitment, we draw attention to the fact that the choosing of reliable partners involves both a reworking of the agent's motivations and a means by which we can recognise this in others. The distinction between the motivation-altering and information-carrying aspects of commitment will become more salient in cases where commitment is not based on shared activity and senders and receivers occupy different roles. For example, in the case of a linguistic promise, if an agent promises to collect figs in a foraging excursion, this is motivation-altering for her, but it is not motivation-altering for the receiver of the commitment signal. Rather, it changes her expectations of the committed agent's cooperation and this is what makes cooperation with her a profitable venture. Now that we have understood how participation in group hunting might be considered a commitment under

Sterelny's (2012) view, I will present the case that participation in group hunting is a commitment under the alternative game-theoretic operationalisation given in Chapter 1.

2.4 Committing to hunting clarified

Recall our earlier definition: *a commitment is a pre-play signal in a strategic interaction taken at time t, that increases the sender's relative payoff for carrying through option X at time t+n, and increases the receiver's probability of the sender carrying through option X.* The subtype of commitment we are working with is an unconditional promise. Here, the move, or commitment signal, the agent takes at time *t* is going on the hunting excursion or showing up to the hunt. The option *X* the agent is committed to carrying through is pulling one's weight in the hunt. One might ask why the commitment signal is not itself pulling one's weight in the hunt. To see why this is so, we must pay heed to the fact that to go on the hunting excursion does not guarantee that the agent pulls his weight when the prey appears. He may instead lag behind and attempt to gain the fruits of the hunt without active involvement, reducing his own costs. However, when we consider that he may be excluded from the profits of the hunt should he go on the hunting excursion and lag behind, we see that to participate at all is to ensure that he has driven up the costs of defection. It is to undertake a pre-play commitment to an execution move that is potentially optimal for him in the long run. It is potentially optimal in the long run because following through on one's commitment reveals trustworthiness. This trustworthiness then allows the agent access to the benefits of future cooperative interaction. Provided these benefits are great enough relative to the gain of defection on any particular occasion, commitment will secure cooperation.

Under this understanding, hunting would operate by taking the "commit to Stag" branch in the extensive form Stag Hunt in Figure 2, where the pre-play signal "commit to Stag" takes the form of going on the hunting excursion.[39] In the simplified evolutionary model presented in the previous chapter, trustworthiness is captured in the sender's likelihood of playing Stag and the receiver's perceived trustworthiness of the sender is captured in the receiver's expectation that the sender plays Stag, $P_C(R1)$. Commitment creates social bonds which open up opportunities for future beneficial interaction since the strategy *SR1 & R1* increases $P_C(R1)$, incentivising *C1* on the part of the receiver since this yields a higher payoff. Observation of cooperation in past plays will increase $P_C(R1)$ in subsequent rounds.

As mentioned in Chapter 1, initially choosing to go on the hunting excursion might first be understood as a cue rather than a signal. That is, it is a trait which does not evolve for the purpose of communication, yet it does offer information that the receiver of the cue can use to condition their response on. The transition looks as follows. If showing up to the hunting location increases the likelihood of cooperative behaviour on the part of other individuals and results in a successful hunt, the sender's predisposition to go on the hunting excursion is reinforced. The sender could then choose to manipulate the behaviour of the receiver because they can anticipate their response to the sender's behaviour – they can choose to convey information about their intended participation or non-participation in a group hunt by showing up or not. This is what Barrett and Skyrms (2017) refer to as ritualisation. As such, behaviours which were initially revelatory about a sender's fixed disposition could evolve to be signals to communicate future behaviour.

[39] Of course, properly understood this should be an *n*-player Stag Hunt game or a Threshold Public Goods game, but we use the two-player representation in Chapter 1 for conceptual clarity.

We now know what commitment via shared activity involves more precisely and why hunting appears to fit this model rather than the costly signalling model. We must also ask whether real life hunting practices fall well into this framework. While we do not have direct evidence of hunting practices from our ancestral environment, we do have ample evidence from modern forager societies. Sterelny (2012) argues that modern data on foraging societies represents a "conservative test" of ancestral conditions. That is, "the ancient-to-modern would tend to dampen down a class of important features of ancient forager lifeways, ones that make cooperation more important. So if we still find those features playing a role in the lives of modern foragers, we can reasonably project them back onto the lives of ancient foragers" (Sterelny 2012: 90).[40]

In order to fully substantiate the thesis that going on the hunting excursion is a commitment, we would want to know: (a) there are opportunities for future interaction, else exclusion may not be detrimental, and (b) that defection following a commitment is punished by exclusion. With regard to (a), hunter-gatherer communities are intimately connected, with reliance on others for food and collective care of offspring, meaning there would be many opportunities for beneficial interaction. There is no direct evidence of hunting *generating* such opportunities, as Sterelny (2012) suggests, but it is clear that many other closely related cooperative activities in hunter-gatherer societies do.

Consider, for example, the ethnographic literature on contingency in food sharing (Kaplan & Gurven 2005; Gurven 2004). Among the Pintupi community, large game is distributed to members of the residential group who have shared with hunters in the past (Myers 1988). A member of the Marmaindê community stated "if one doesn't give, one doesn't get in return… some people are

[40] Of course, a difference to note is that modern foragers possess material technology that our ancestors would not have had (Marlowe 2005).

specifically excluded from most distributions because they never or only rarely give any of their products to us" (Aspelin 1979: 317). Indeed, Aspelin suggests the exchange value of food in tying people together may be greater than its use value. He writes that "Marmaindê food distributions are thus both symbolically and practically one of the critical points in the social cohesion of their social world" (Aspelin 1979: 325).[41] These pieces of evidence show that, in at least some forager societies, future benefits are contingent upon current cooperation, so what explains current cooperative behaviour is the fitness consequences to be gained from future opportunities for beneficial interaction. It is likely that commitments to hunting open up such possibilities by establishing an agent as a keen provisioner of food in the social network.

With regard to (b), one might wonder whether exclusion itself faces a free-rider problem. That is, why would an agent punish free-riders if there are others in the group who could do so, and she may free-ride off their punishment of free-riders? To punish imposes a cost on an agent but benefits the group, so punishment is itself an altruistic behaviour in need of explanation. This is known as the second-order problem of altruism (Boyd et al. 2005). Exclusion from the spoils of the hunt is indeed difficult to explain since it involves excluding agents from a public good, which may involve costly activities such as driving free-riders away. However, for showing up for the group hunt to count as a commitment, it is sufficient that agents are excluded from *future interaction* upon defection. Punishment in the form of exclusion from the spoils and from future interaction of those who do not produce or share enough food in hunter-gatherer societies is widely documented (Gurven 2004).[42]

[41] Despite the evidence presented here, contingency in food sharing is not observed in all forager societies. See, for example, the sharing practices of the Ache with respect to large game (Kaplan et al. 2005).
[42] The production of food here is not limited to hunting activity and may also include gathering.

Gurven describes collusion between two Guwinggu family clusters to share less food with a third cluster who was not producing enough. The sanction induced higher production and sharing by the third cluster. Agta civilians who are unproductive are socially ostracised until they are forced to relocate (Griffin 1982). Most relevant to the current thesis, it was found in the Netsilik Eskimo community that "*lazy hunters* were barely tolerated by the community. They were the objects of back-biting and ostracism… until the opportunity came for an open quarrel" (Balikci 1970: 177, my emphasis). Importantly, Gurven (2004) also argues that success rates for hunting game depends on luck and random factors, making it likely that defectors are punished for not contributing enough *labour* rather than for not producing enough food. These observations accord with the suggestion that lagging behind in the hunt is subject to potential exclusion from the benefits of future interaction. It thus supports my claims regarding hunting being a commitment since defection from hunting demonstrably entails punishment.[43] Signalling one's intended participation increases the cost of defection. If this is so, commitment offers a good explanation for the stability of group hunting in the face of short-term incentives to cheat.

One might object that the commitment signal is not doing any work here – the exclusion of free-riders by preferential interaction is the mechanism sustaining cooperation. However, as discussed in Chapter 1, at some level of description, exclusion of defectors is doing the work in *any* account of the evolution of cooperation that rests on correlated interaction, since this is simply what

[43] There is also some evidence of tolerance for slacking and tolerated theft. Among the Siriono, the elderly are known to steal food. Though not directly punished, they are subject to negative gossip (Holmberg 1969). Endicott (1988) also describes situations among the Batek in which adults are not expelled for taking more food than they produce. This does not wholly undermine my thesis, however, as we would expect a stable level of defection to coexist with cooperation in equilibrium.

correlated interaction amounts to. The theories differ in what mechanisms lead to correlated interaction. The mechanism in this account is the signal of intended participation. It is those who signal commitment by showing up to the hunt then do not pull their weight that are excluded, since it is going on the hunting excursion that changes receiver expectations of cooperation. This is a different explanation of cooperation than one based on, for example, correlated interaction due to proximity of cooperators, though both ultimately amount to the exclusion of defectors (Skyrms 2004).

Why believe that the signal of participation is the mechanism behind correlated interaction? This framework accords with what we see in practice. Women and children who stay at the camp have not signalled participation and so are not renegers if they do not hunt. In accord with the theory of commitment, they still receive a share of the food, since it was never expected that they hunt. Sharing to non-participants in the hunt is evidenced in the Hadza, Dobe !Kung, G/wi, Ifaluk, Ache, Yanomamo, and Gunwinggu societies (Kaplan & Gurven 2005). Non-hunters who do not go on the hunt are not, properly understood, free-riding. They are often providing a reciprocal service in exchange for shared food. Indeed, Gurven (2004) documents that trade plays a major role in sustaining food sharing. Pintupi women give food to those who look after their children, Yuqui and Tsimane hunters trade their kills for garden products, manioc is given to Kuikuyu in exchange for weeding labour, and Siriono and Yanomamo men give food in exchange for sex.

Another important case is children who go on the hunt but do not contribute to capture of the prey, instead learning via observation (Puri 2005). They are not excluded from the benefits of cooperation so one might think this provides a counter-example to the thesis that showing up to

the hunting excursion is what makes one vulnerable to exclusion. However, we would not want to say that a child showing up to the hunt constitutes a credible commitment to participate. Other observable characteristics of an agent, for example, age and physical well-being, can render their presence an insufficient signal for the receiver to expect cooperation. If he does not, the receiver will not exclude the agent based on insufficient contribution. As such, the sender's relative payoffs for the future action remain unchanged, meaning no commitment has been made.

What about the *able-bodied man* who stayed at the camp instead of going on the hunt? We might suppose that he is excluded from cooperative benefits yet it might not appear that he has signalled. All that matters is that able-bodied men should hunt and they are punished if not. If this is correct, then there appears to be a simpler explanation of group hunting that does not rely upon commitment. I am sceptical of this hypothesis. He would pay a cost if he was excluded from the spoils of the hunt or other cooperative activities in the future. This could only be so if others have an expectation that he hunt as a result of (a) norms of behaviour that are attached to a particular subgroup of the population, perhaps inferred from observable characteristics of the individual in conjunction with observed behaviour of relevantly similar individuals, or; (b) past instances of behaviour which served as signals. In (b), exclusion follows a signal that entails fitness consequences; it follows from a prior commitment.

The explanation is unlikely to be norms, as in (a). Shared social norms related to hunting which were nuanced enough to demarcate relevant subgroups of a population based on their characteristics and to exclude them based on expected behaviour appear quite sophisticated. While there could have been more minimal social norms at play approximately 2 mya, for example, in

kinship organisation, it is likely that these more sophisticated social norms occurred closer to the evolution of cultural groups. Though there is debate surrounding the evolution of cultural groups, some place the first clear signs as late as with *Homo sapiens sapiens*, at 200 kya (Tomasello 2014). My position here turns on an empirical claim which has not yet been sufficiently studied but, if we suppose that such sophisticated norms evolved closer to the evolution of cultural groups, the intuition that the able-bodied hunter will be excluded is based on our current norm-laden psychology and may not have been the case in the ancestral environment in the absence of commitment.

Perhaps sex-based roles, rather than full-blown cultural norms, are all that is required to explain the supposed exclusion of men who go on the hunting excursion and lag behind as well as those men who do not go on the hunting excursion at all. That is, if men are expected to hunt and women are not expected to hunt, able-bodied men who stayed at the camp and did not signal their commitment would still be punished for not committing, undermining the fact that the commitment signal is important. There are sex-based differences in behaviour in many species of animals. A prominent example is that of chimpanzees, whose adult males take on the role of patrolling territory boundaries to deter or kill members of other chimpanzee communities. However, despite these clear sex-based differences in behaviour among chimpanzees, there is no evidence of punishment for free-riding and violating expectations (Watts & Mitani 2001; Wrangham 1999). Indeed, this is dependent not just on differences in what males and females *do* in the community but norms about what they *ought* to do and what punishment they deserve if they fail to do so.

Therefore, the evidence needed to substantiate the claim that sex-based differences in behaviour suffice to explain the stability of group hunting would be that male cooperatively hunting species punish males who do not signal they will participate in the hunt. There is as yet no evidence of this in hominins or closely related species. However, we do have evidence that those who signal participation and *subsequently* defect are punished, in support of the commitment hypothesis. Perhaps this evidence could be equally well explained by sex-based expectations if the expectation is that men are *effortful* hunters so they would be punished for both lagging behind after signalling participation and for staying at the camp. However, I reject the idea that such sex-based expectations offer a *simpler* explanation of the stability of group hunting than commitment. Both explanations turn on an actor being able to condition their response on their expectations and so require sensitivity of dispositions to the behaviours of others. Whether these expectations are a result of observable signals (as in the commitment hypothesis) or observable characteristics (as in the sex-based differences hypothesis) of an agent will not matter for the cognitive or social complexity involved in the explanation.

Yet, if signalling participation increases the cost of defection, why signal at all? That is, if agents who did not show up to the hunting excursion and therefore did not signal their intended participation were *not* excluded from the spoils or future interaction, we ought to question why agents would ever choose to go on the hunting excursion. I argue there is something to be gained from going on the hunting excursion. Those who go on the hunt hope to receive a larger share of the game. In a comprehensive survey of food-sharing by Gurven (2004) on 33 hunter-gatherer societies and 13 forager-agriculturalist societies, it was found that producers often receive significantly more than a proportionate share of the food, thereby making the hunting of large

game worthwhile.[44] Examples of hunters pooling catches among themselves are found among the Mbuti, Aka, Washo, Hiwi, Pintupi, Northwest California Indians, Netsilik Eskimo, Inujjuamiut, Fanalei, and Maori. Indeed, hunters are rarely criticised for consuming organs and marrow at the kill site. This is seen in the !Kung, the G/wi, the Nyae Nyae !Kung, the Hadza, and the Batek. Kaplan and Gurven (2005) also find that Yora hunters first take a portion of the food themselves and then distribute it to other families. The products of cooperative fishing and whaling among the Ifaluk, Lamalera, and Makah are similarly first distributed to participants in the activity, and then distributed to non-participants in the group. This benefit explains not only why cooperators go on the hunt but also defectors. The defecting agent who goes on the hunt and lags behind but yet gains access to the kill fares best, since he receives a greater share of the game but exerts little to no effort. It is this temptation to cheat which motivates the defector to signal.

A cautionary note: despite this evidence of punishment by exclusion, we will want to be careful about making inferences from facts concerning modern hunter-gatherer societies to information about our early hominin ancestors.[45] However, we do not need this ethnographic literature to provide conclusive evidence of the behaviours of our ancestors. It constitutes our best available information but even if this were questionable, in reality, we need very little to get commitment off the ground in hunting. All that is needed is that those who signal cooperation and defect are *less preferentially chosen* in future interactions. Punishment on this account amounts to partner choice and this has a deep history.

[44] While, in modern hunter-gatherer societies, such distributions are often based on social norms, this does not preclude hunters receiving a greater share at the kill site even in the absence of such norms.
[45] The issue of such "ethnographic analogies" from the present to past is discussed in detail in Wylie (1985) and Currie (2016).

The cognitive prerequisites involved in recognition and preferential interaction are present among non-human primates.[46] Suchak et al. (2016) conducted a study where a joint task required two or three chimpanzees to work together toward a larger food reward than they could attain individually. Crucially, the experiment allowed a group of chimpanzees to perform the tasks alongside one another, involving observation of others' actions and free partner choice without experimenter intervention. Suchak and colleagues suggest the tendency to seek cooperation with similarly ranked partners may have helped deter free-riding in these experiments, since chimpanzees with a dominance advantage were more likely to engage in free-riding. Chimpanzees who were victims of free-riding also withdrew from the task 24.6 per cent of the time or avoided engaging until the free-rider moved away 20 per cent of the time. By the end of the experiment, the chimpanzees cooperated "almost non-stop", suggesting that partner choice is an important mechanism for securing cooperation (Suchak et al. 2016: 10218).[47] If the capacity for preferential interaction is present in non-human primates, this was also likely present in early hominins. As long as maintaining interaction is important for hunter-gatherers, partner choice serves as a low-cost form of punishment underwriting commitment, since potential exclusion raises the cost of defection. One might worry that hunters could not be excluded from the hunting party in the future, but it is sufficient that they are excluded from other cooperative interactions, such as food-sharing, which we have seen evidence for.

The proximate mechanisms by which commitments are carried through are emotions as well as the more specific kind of "arousal" exemplified in cases of combat-like shared experience suggested by Sterelny (2012). On Frank's account, emotions cause us to reevaluate our potential

[46] See also Schino and Aureli (2009) and Bshary and Grutter (2002) on partner choice in animals.
[47] Alongside direct deterrence and third-party punishment.

options at the commitment trigger such that the subjective payoff of the option to which we have committed is higher. The agents who can credibly commit to cooperation through emotional mechanisms often end up with better material gains since it is these agents who are trusted and chosen in future interactions. For example, the agent who commits to continuing a marriage out of love will secure the material gains of her partner's cooperation as well as signal her virtuousness to others. This fitness benefit reinforces the evolution of motivationally powerful emotions. In the case of hunting, the emotion involved might take the form of loyalty or positive affiliations towards one's hunting partners.

Furthermore, as Sterelny suggests, we see evidence that emotions are amplified and take on a different character in situations of combat. It has been found that drill plays a crucial role in maintaining group cohesion in war, shifting the individual sense of fear toward a collective experience of bonding (Malešević 2021; Holmes 1985). McNeill (1997: 2) writes of his own experience in World War II that drill and unified marching generates a strong "muscular bond" resulting in an emotional change: "a sense of pervasive well-being is what I recall; more specifically, a strange sense of personal enlargement; a sort of swelling out, becoming bigger than life, that's to participation in collective ritual."[48] So there is also real life evidence of the proximate mechanisms Sterelny suggests are at play in this form of commitment.

Of course, I present the above arguments not to say that commitment was the *only* factor in sustaining cooperative hunting. Kin selection, for example, will have also played a role. However, as mentioned in Chapter 1, these theories are not mutually exclusive. Both Trivers (1971) and

[48] See Malešević (2021) for an overview of emotional responses to war.

Wilkinson (1988) have pointed out that reciprocity can occur between kin and "generate a substantial selective force independent of kin selection even when performed among related animals" (Wilkinson 1988: 98).[49] Furthermore, hunting is not only informationally demanding in the short-term as a result of the demands of coordinating cooperative interaction, but it is informationally demanding in the long-term because it requires transmitted skills and expertise (Sterelny 2012). As such, longer-term social learning and culture were also crucial to its stability. Tracks and other physical signs of passage provide important information and much of the information concerning tracking is acquired culturally. For example, Aboriginal children learn how to recognise tracks by being shown how to reproduce them (Sterelny 2012). Adult actors within a group will leak information in everyday activities as well as adaptively structure the learning environment of young in order to prepare them for skill acquisition. This might concern tool- and weapon-making, as well as preparation of potentially poisonous foods to make them safe for consumption. Indeed, by 400 kya, humans were using material technology that could not have been reinvented each generation, so was likely passed on through cultural transmission (Sterelny 2012: 14). Commitment is thus one of the mechanisms which lessens the short-term coordination difficulties of hunting. It both incentivises sender cooperation and transmits information about reliability to receivers. Yet commitment will work in tandem with social learning and cultural transmission to ensure the long-term stability of cooperative hunting.

So I argue that one of the major cooperation-sustaining mechanisms behind group hunting was commitment. Hunting is a stable form of cooperation because to go on the hunting excursion

[49] While the more general explanation provided by reciprocal altruism suffices to explain cooperative hunting, the more detailed explanation of the evolution of cooperation provided by the operation of commitment is illuminating. See Chapter 1 for discussion.

incentivises an agent to pull his weight in the hunt. This is because subsequent defection will result in potential exclusion. A commitment can make group hunting more robust by adding an additional cost to defection. If an agent reneges on hunting activity in the absence of a commitment, it is natural to suppose that he will not receive a share of the spoils. However, with a commitment in place, reneging not only jeopardises access to the spoils or reciprocal care, but also fractures social bonds and reveals untrustworthiness, leading the agent to be excluded from the potential benefit of future interactions, too. As such, commitment raises the cost of defection, making cooperative hunting more easily sustained. Indeed, once cooperation is widespread, the possibility of exclusion from future interaction is likely to outweigh the benefits of defecting on any particular occasion. At the same time, the commitment allows other members of the hunt to identify the agent as trustworthy and incentivises receiver cooperation. This account can be contrasted with the costly signalling view, which sees hunting as a signal of phenotype quality. I argue that hunting transmits information of a different sort to recipients – it transmits information about the intended action of a sender.

2.5 On deception and exclusion

In this section, I will address the problem of dishonest signals of commitment and whether this undermines the signalling system. I argue that it does not. First, I will outline two reasons why we might think there arises no problem of deception – owing to Frank and Sterelny – and I will argue that these reasons do not apply to commitment via shared activity. So, there is a problem of deceptive signalling that needs to be addressed. Second, I will establish that social costs can keep commitment signals honest in the face of potential deception. Third, I offer some suggestions for

why we might be incentivised to engage in costly punishment in excluding others from a public good, though such costly punishment is not strictly required for the commitment account.

On Frank's (1988) account of commitment mechanisms, deception is controlled because the person who wishes to deceive others about her emotional state must undergo cognitively demanding executive management to make her emotions seem credible, whereas the person who genuinely acts on emotions does not. This cost asymmetry between honest and deceptive signals is built-in and is what serves to make commitment signals credible to receivers. Here, Frank appeals to the costly signalling framework to account for the credibility of commitment signals. Sterelny (2012) also notes credibility for a dishonest agent is not cost free, since a dishonest agent must gather the evidence to demonstrate the plausibility of her claim. For Sterelny, though, this does not depend on an in-built cost asymmetry between an honest and deceptive agent. Rather, "the difference in relative cost is driven by the general fact that faking a property always depends on signal production. It is never an informative by-product of the other behavior, a by-product that can be co-opted as a cheap signal or as evidence that makes a signal credible" (Sterelny 2012: 112).

Sterelny (2021a) also argues there is little problem of deception in hunting since hunting is an instance of *immediate return mutualism*. Return mutualism concerns those cooperative endeavours in which the benefits of cooperation are received immediately upon successful completion of the activity. This is contrasted with reciprocation which may involve benefits conferred at a later date or of a different nature. Sterelny claims that in cases of immediate return mutualism, if an agent cheats, defection is clear to participants and all participants in the hunt are harmed agents, meaning

that each is individually incentivised to enact punishment on the cheater. So, he argues, deterrence of cheating in small-scale collaborative activities does not face the same difficulties as deterrence in reciprocal exchange.

I do not believe either of these suggestions offer us a plausible way out of the problem of deception. Participation in shared activity will not be made honest by intrinsic costs as in Frank's account. It is not obvious that faking participation in a shared activity is too costly for the agent. Leaving aside psychological constraints, it is plausible one feigns active participation in hunting but effortlessly lags behind. So, though emotion may result in an intrinsic cost asymmetry in some of the commitments I am concerned with, my account will not turn on this. Furthermore, Sterelny's assertion that defection is clear in cases of immediate return mutualism does not apply to all hunting activity. Sterelny's claim only applies to one of two separate games present in group hunting. One is an *n*-player Stag Hunt where the agents capture prey on the hunting excursion and the other is a Bargaining game where they divide the spoils of the hunt. Monopolisation of the prey is a form of defection in the Bargaining game which is obvious to all. However, defection in the case of hunting activity is not necessarily obvious. If hunting is fast-paced and involves surrounding prey from multiple directions, it may be difficult to tell whether an individual is pulling his weight. Indeed Pickering (2013) and Bunn and Pickering (2010) argue that ambush hunting is likely the first form of hunting, occurring earlier than endurance or persistence hunting. This is a form of hunting in which it is even more difficult to discern contributions due to distance. So $P_R(d)$ in the case of hunting activity may be much lower. If this is so, we cannot appeal to immediate return mutualism to explain the deterrence of cheaters.

Instead, I argue that the credibility of commitment signals can be ensured by social costs. If defection from joint activity in small-scale collaboration is directly observed and punished, and if the punishment of false signalling exceeds the benefits of defection, dishonest signals will not be stable. Recall the payoff of defection from commitment signalling presented in Chapter 1: $EV(SR1 \ \& \ R2) = P_R(C1)x + P_R(C2)y - P_R(d)c$. In this equation, even if the perceived probability of one's defection being detected, $P_R(d)$ is low, defection will be disincentivised as long as c is sufficiently high. What does it mean for c to be high? This means that the value of current or future interaction is worthwhile, so exclusion imposes a significant cost. Indeed, in hunter-gatherer societies, c is likely to be very high given that agents rely upon one another for provisioning, collective defence, and alloparental care. The cost of exclusion from these activities could be so great as to be a threat to survival. Indeed, lone hunters would likely not be able to provide for their family and, as we have seen earlier, much of hunter-gatherer provision and care is reciprocal. Thus, to risk exclusion could be devastating. Here, the relevant punishment is exclusion from cooperative interaction. So social, if not intrinsic, costs can serve to keep such signalling systems credible in small-scale interactions even where the probability of detecting defection is low. Of course, emotions such as anger likely play a proximate role in incentivising punishment (Hirshleifer 1984), even if we do not take emotions to render deception intrinsically implausible, as in Frank's account.

Given that we are relying on social cost to maintain honesty, one might wonder again about the second-order problem of altruism (Boyd et al. 2005). However, as noted earlier in the chapter, costly punishment in the form of exclusion from a public good is not strictly required to render showing up to the hunt a commitment. Punishment on this account need only amount to preferential partner choice in future interactions, which is demonstrated in animals and directly

benefits the excluder since she avoids interacting with an agent who she believes will sucker her. One might wonder whether such exclusion means the excluding agent has no way to survive if survival depends on cooperative tasks. However, exclusion is not an all-or-nothing matter but a matter of relative attractiveness. Exclusion involves *ranking* potential partners according to their suspected cooperativeness in a strategic interaction. If no other partner is available, it may pay to opt for a less preferred partner and risk exploitation, if the pull of potential success is great enough.

Yet the account would be stronger if we could substantiate the claim that agents are also excluded from current interaction where the profits of current interaction take the form of a public good. Henrich and Boyd (2001) develop an evolutionary model in which second-order altruistic punishment can be sustained by an arbitrarily small amount of conformist transmission, where conformist transmission is the tendency to copy the majority. Ozono and colleagues (2016) show how, in public goods games, leaders in a group can stabilise cooperation by punishing both non-cooperators and non-supporters (those who do not punish other non-cooperators) and that leaders in fact benefit when this is the case, favouring the evolution of punishment and second-order punishment. Sasaki and Uchida (2013) also note that exclusion involves a natural benefit to the agent because exclusion decreases the number of beneficiaries of a public good. As long as the cost of driving the free-rider away is less than the reallocated benefit, exclusion will be incentivised.

It is possible that one or more of these explanations apply to hunters. It is plausible that there is selective pressure for conformist tendencies in bands where signalling group membership is important for survival. Indeed, even chimpanzees exhibit conformist tendencies within the group.

A study concerning the tool preferences of three different (but genetically related) chimpanzees in the Taï National Park found evidence of different, purportedly cultural, practices among the groups (Luncz et al. 2012). Coula nuts dry out over the nut season and become progressively easier to crack. If chimpanzees use the most efficient tools, we should expect that all communities use the rarer but more effective stone hammers at the beginning of the season to open fresh nuts and the less powerful but more common wood hammers at the end of the season when nuts are drier and thus easier to break. The study by Luncz and colleagues (2012) found that the three groups differed in which tools they used and the size of hammer used. Furthermore, when a female migrated, she adopted the tool use of the new community, suggesting the tool preferences of the group are rules that are socially acquired through a conformist tendency. Though there is no direct evidence of the conformism of our hominin ancestors, it is likely conformism is a plausible explanation of our ability to overcome the second-order problem of altruism and punish free-riders.

Alternatively, a natural ordering of strongest to weakest individual in the hunt may provide a less formalised analogue to Ozono and colleagues' (2016) leadership role. It is possible that weaker individuals in the hunt are not able to drive off defectors but a stronger individual has the power to corral support. This explanation would rely on individuals recognising their relative ranking in the hierarchy, but this is not an implausible assumption. There is ample evidence that chimpanzees keep track of dominance rankings and use these to make assessments about interactions with potential partners. However, it also relies on the defector not being the most powerful member of the group which is less likely. It is also worth noting that, if commitment evolved first in dyadic interactions, the second-order problem of altruism may not be so troubling. In dyads, there is no possibility of free-riding off the punishment deployed by others, and the exclusion is itself better

for the agent doing the excluding since she avoids future interaction with an exploitative other. The proximate mechanisms we developed to incentivise fitness-enhancing exclusion in these cases – notably, anger – might have continued to operate in the group context.

In sum, if hunting activities are spatially distributed and fast-paced, cheater detection may not be so easy. I argue that commitment signals are not primarily kept honest through intrinsic costs, but rather social cost. This takes the form of exclusion from the benefits of current interaction as well as potential future interaction. If individuals interact in small and frequent ways, which is likely if their interaction was important for survival, the costs of defection are compounded. Though exclusion from future activity is all that is required for showing up to the group hunt to count as a commitment, I have offered some reasons to believe that exclusion from current interaction, which concerns exclusion from a public good, is possible among our hominin ancestors. So, I argue the problem of deception did not undermine commitment signalling in the case of shared activity. As we will see in the next chapter, deception becomes a much more pressing problem in larger groups with the capacity for temporally extended commitments and many opportunities for new partnering.

2.6 The emergence of larger, multi-level societies

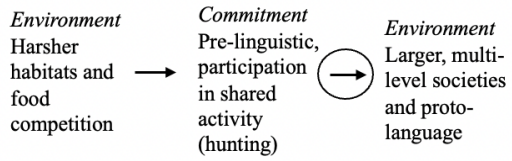

Figure 8: The emergence of larger, multi-level societies.

In what follows, I elucidate how commitment via shared activity provides the selective environment for more cooperation to evolve. In particular, I argue that small-scale collaboration based on commitment played a role in growth of the group size or social network. In Chapter 3, I show how the formation of larger social networks and the requirement to share information in joint tasks laid the groundwork for the evolution of language. Linguistic information-sharing is a form of cooperation that coevolves with commitment based on shared activity, and which underpins the evolution of our second form of commitment, linguistic commitment.

Sterelny (2012) argues that the cooperation seen in early hunter-gatherer societies permitted the hominin lineage to expand into new habitats and reduced the risk of predation from other species. Tasks that surrounded the practice of group-hunting, for example, weapon-making and food preparation crucially relied upon cumulative cultural learning. As our capacities for effective cultural learning developed, we were able to extend this skill to innovations in new spheres: the development of clothing, shelter, and even more sophisticated tools. Such developments contributed to an extended lifespan due to better protection against harsh habitats. Division of labour in hunting and gathering provides a resource buffer when prey is not available. This increased means of subsistence allows the group to grow and specialisation in these skills allow it to become more productive. Specialisation also increases scope for innovation. As a result, we may see a positive feedback loop between collaborative hunting, increased group size and the rate of innovation (Sterelny 2012).[50]

[50] Of course, this is not to say that innovation increased at a consistent rate following these developments. There are many periods of debated stasis in the early Acheulean.

Foley and Gamble (2009) argue for several key social transitions in human evolution, many of them relating to the value of food resources. One of these key transitions occurs at 2.6-1.6 mya with the advent of tool-making and meat consumption. Greater access to animal resources and smoothing of the food supply across seasonal variation led to greater availability of energy for mothers, a relaxation of the energy constraints on development of larger brains, and elongated life histories (Foley & Gamble 2009; Foley & Lee 1991). Furthermore, it is supposed that this led to more prolonged male-female bonding in early *Homo,* an initial form of cooperation involving greater male investment in offspring and a divergence of male and female roles. The next key transition occurs with the development of fire and cooking of more energy-rich food sources around 800-700 kya with *Homo heidelbergensis*. Foley and Gamble argue control of fire would have had a major impact on interdependence of individuals and the role of males and females in subsistence, which were contributing factors in the establishment of nested family units within the community structure. Thus, we can see that hunting and the change in food preparation that resulted was key in the evolution of cooperation. All of these aforementioned changes in social structure may have contributed to growth of the group size.

Unlike non-human primates, hunter-gatherers exhibit a hierarchical social structure of co-residing families with friendship dyads across camps, frequently referred to as the multi-level society. Since residential camp size in the modern ethnographic literature shows groups to be between 30 to 50 individuals, growth was unlikely to be at this level, but rather in the development of the interconnected network of camps called the village, comprising of approximately 100 to 200 people. Modern hunter-gatherer societies also have larger groups called tribes, of roughly 500 to 2,500 individuals (Dunbar 1993). While Migliano and colleagues (2020) suggests this transition

to multi-level organisation emerged around 320 kya, Layton and colleagues (2012) suggest this occurred much earlier, at 800-700 kya with *Homo heidelbergensis*.[51] The authors use evidence such as neocortex ratios, the distance materials had travelled from their source, ethnographic data on hunter-gatherer daily foraging ranges, population densities, and fossil hominin morphology to substantiate their thesis.[52] They argue that the larger village emerged as a result of the shift to meat-eating with bipedalism and hunting, and the effect this had on population density. In particular, early hominins consumed 10 times more meat than chimpanzees in the forest and this greater reliance on meat encouraged increased dispersion. Layton and colleagues (2012) argue that meat also contributed to encephalisation since it is easier to digest, leaving more energy for the brain. The authors further cite cooperative breeding as a contributing factor in hominin encephalisation. This is important since it is thought that group size is constrained by neocortical volume – the larger the neocortex, the more individuals an agent is able to successfully keep track of (Dunbar 1993, 1996).

The increased foraging distance necessitated by reliance on predation meant that individuals became so dispersed that they could no longer encounter all members of the group during their daily interaction. Residential camp formation with an overlapping social network then occurs by a process of expansion and fracturing of the initial group (Layton et al. 2012). In particular, formation of distinct residential camps become adaptive when the daily foraging range cannot

[51] Grove et al. (2012: 195) suggest even earlier, arguing that hominin species from Australopithecus afarensis onwards "would have periodically divided their groups into at least two subgroups." In their view, early Homo, *H. erectus*, and the "*Homo*-like" Australopithecus sediba would have fissioned into three subgroups shortly after 2 mya.

[52] For example, in the *Heidelbergensis* community at happisburgh, 15 per cent of Lower Paleolithic depositions occur more than 25 kilometres from their source, with the furthest definite case transported 65 kilometres. This suggests that *Homo heidelbergensis* was in an environment where the nested band formation would have advantageous (Layton et al. 2012).

extend across the entire area occupied by the individual's social network. In this environment, the unpredictable occurrence of large game is best exploited by sharing at residential camps. Since, with expansion in the size and area of the group, it is more difficult for the community to reform, what were ad hoc foraging groups became the stable locus of daily interaction, nested in a village community of which they were once part. These village communities offer a number of other advantages that help explain their emergence, for example, mutual assurance against environmental disaster such as drought and the resolution of disputes by migration (Layton & O'Hara 2010). Such village ties may kept stable as a result of broader kinship recognition since the advent of pair-bonding (Chapais 2008; Sterelny 2021b).

One might be sceptical about the particularities of Layton and colleagues' account. However, all we need for my overall argument is that small-scale collaboration based on commitment has some role to play in growth of the group size or social network, since it is in the context of these larger groups that we see the evolution of language. The model of Layton and colleagues has merits. In particular, we trade the more difficult explanation of how distinct camps came to know one another and become cooperative with the simpler problem of how an initial group maintains trust as it fragments into smaller groups (Sterelny 2021b). However, we do not need to subscribe to this model in order to explain the role of commitment via shared activity in the emergence of multi-level societies.

Alternative hypotheses include the ancestral male kin group hypothesis (Chapais 2008) and the dispersed resource hypothesis (Aureli et al. 2008; Foley & Gamble 2009). In the former, it is thought that multi-family units are a derived trait from the last common ancestor in their likely

development of polygynandrous mating systems, female dispersal, male philopatry, and male kinship bonds. In the latter, the transition to more open savannah habitats made grouping advantageous as a defence against predation risk. This was alongside the increasing importance of underground storage organs to the hominin diet, such as roots and tubers. Together, these environmental conditions would have compromised the spatial cohesiveness of the group and favoured substructuring. Despite these other potential factors, it remains highly plausible that group hunting (and perhaps other shared activities based on commitment) contributed to growth of the group size given that it allowed a smoothing of the food supply, greater energy for mothers, and the growth of expertise and innovation following a division of labour.

Not only did group hunting contribute to growth of the group in the form of a multi-level society, but the requirement to share information in joint tasks such as group hunting and cooperative breeding provides us with the cognitive prerequisites to language, argued in more detail in Chapter 3. In particular, selection favoured the evolution of certain perspectival representational capacities, socially recursive inferences and self-monitoring abilities in order to effectively engage in such communication. Importantly, the multi-level society constitutes an environment in which we are potentially required to engage in cooperative interactions with new partners on novel tasks. This is a situation in which we do not have a shared common ground, creating selective pressure for the development of a group-wide means of communication. This kind of conventional linguistic communication requires the development of certain cognitive capacities and these are importantly continuous with the capacities developed in our protolanguage. In particular, in order to effectively communicate with in-group strangers, it is necessary that agents are able to adopt an agent-neutral perspective or schematise certain situations in an agent-neutral way, using a culturally-constructed

representational system, and to govern their thinking and communication by the normative standards of the group (Tomasello 2014).

Another important development in early hominin evolution was the advent of cooperative breeding where childcare is provided by adults in the group other than a child's parents. The fossil records cannot tell us exactly when cooperative breeding occurred in our evolutionary history, but possibly as early as 1.8 mya (Hrdy 2009). Cooperative breeding also resulted in an expansion in terrestrial habitats as well as the development of new cognitive capacities geared toward cooperation (Kingma 2017; Hrdy 2009). The developmental environment of an infant was richer in virtue of interactions with individuals outside of the natal group, including males and other juveniles (Sterelny 2012). This is a drastic change from the infants of primate species who are not reproductively cooperative. Here, infanticide is common and mothers are highly protective of their young (Hrdy 2009). If there are a greater number of information sources for a child, information is transmitted more frequently and can withstand losses to a particular member of the developmental circle. Furthermore, selection would operate on those capacities that make an infant more appealing to potential alloparents and will involve the development of additional cognitive capacities aimed toward interaction with others. In particular, alloparenting will select for infant monitoring of adults, awareness of differences, and awareness of responses to the infant's own actions (Sterelny 2012; Hrdy 2009). These are the building blocks of cooperation. Not only this, but cooperative breeding means mothers can devote energy to producing more babies while others provision their previous young, contributing to growth of the group. Cooperative breeding, by increasing the number of carers, also provides additional security for infants when the group faces harsher conditions, resulting in groups which are more stable (Rubenstein & Lovette 2018).

One can also understand the importance of shared activity in maintaining social cohesion by considering what happens after the need for collective hunting subsides. The "broad-spectrum revolution" involved a transition from large and medium sized game to small and medium sized game and marine resources (Stiner 2001). At the same time, kills were more frequently successful and the need for collaborative foraging and food sharing became less marked (Sterelny 2012). These changes were later intensified by the projectile revolution. Until the Middle Stone Age, perhaps as late as within the last 100 kya, weapons were fairly short-range, likely 10 meters at most (Sterelny 2012: 90). This means that cooperation in hunting was crucial. Lombard and Shea (2021) estimate that humans were not using effective weaponry for distance killing until approximately 90 kya in Ethiopia (with dart use), though there is more substantial evidence of both bow and arrow and dart use in Africa and Eurasia between 90-30 kya.[53] Here, the cooperative landscape changes. Where, previously, the profit of hunting was accrued together, we now have situations where individual success in hunting is possible, though variable. Not only this, but resource incommensurability arises with different food sources, and reciprocation is delayed over time. Here, other forms of cooperation apart from social foraging became more important for the maintenance of group relations, such as cultural norms.[54]

Crucial to our next chapter is that, where shared activity like group hunting is important for survival, selection favours communication and information sharing among members of a group or

[53] See also Shea (2009) on the emergence of projectile technology and Marlowe (2005) for earlier dating.
[54] Indeed, Foley and Gamble (2009) argue that the development of projectile lithic weapons, alongside greater importance of clan and lineage-based kinship systems as well as regional differentiation of communities, introduced a greater cognitive demand for maintaining social relationships over distance and time in *Homo helmei, Homo neanderthalensis* and *Homo sapiens.* They suggest this may have been a "critical pre-condition" for the evolution of cultural capacities that would enforce social norms (Foley & Gamble 2009: 3275).

larger community. It favours the evolution of language. Whilst many explanations of information exchange focus on its role as a means of securing cooperative interaction, this does not preclude that information exchange is itself a form of cooperation. Underpinning this exchange is the coevolution of commitment and cooperation: small-scale cooperative enterprises create an environment which makes the sharing of information adaptively advantageous to the group. While, here, we have focused on the ways in which commitment via shared activity enabled the formation of larger groups, Chapter 3 will deal in depth with the evolution of language and reputation sharing. We will see that early hominins were in a context which favoured the emergence of greater communication as a form of cooperation and that previous practices of commitment had enabled the cognitive foundations for these increased communicative capacities. However, not only does commitment via shared activity aid successful cooperation and expand the realm of possible cooperation, but this cooperation in turn provides the selective environment for more effective forms of *commitment* to evolve. It was against this backdrop of linguistic information sharing that a new form of commitment emerged as a means of altering one's attractiveness as a partner, even where opportunities for shared activity did not arise. We will see that linguistic commitment enables us to make more fine-grained commitments and commitments on a greater variety of issues, facilitating cooperation yet further.

Chapter 3: Linguistic commitment

In the previous chapter, I showed how, in an environment where collaboration was important for survival, commitment via shared activity serves to secure cooperative outcomes. I also discussed the development of larger, multi-level groups as a result of these activities. In this environment, there was growing selective pressure for use of more sophisticated communicative capacities to facilitate cooperation in an increasingly complex social world. In this chapter, I will discuss the emergence of language[55] in more detail. I then draw our attention to the fact that the advent of larger groups and language favoured the development of two new forms of commitment – explicit and implicit promising. These commitments operate in the same way as commitments via shared activity: first, by changing the fitness consequences of the sender since defection entails potential exclusion from current or future beneficial interaction. Second, by changing the expectations of the receiver about the sender's intended actions, aiding in the receiver's discrimination of trustworthy partners from untrustworthy ones.

I argue that linguistic commitment offered a number of fitness advantages over commitment via shared activity. These advantages enhance the effectiveness of cooperative interaction, expand the range of cooperative activities we may safely engage in, and increase the scope of potential partners for cooperation. Alongside these features, linguistic commitment affords agents more opportunities to determine whether a potential partner is trustworthy. So, as the realm of our cooperative enterprises extend, selection favours additional forms of commitment as a means of discriminating among potential partners and making oneself more attractive to others. Linguistic

[55] Though there are differing definitions of "language", for our purposes, I require only that the language we are concerned with is sufficiently rich and flexible to allow one to express statements that concern events at another place and time and to represent value judgements regarding the actions of others.

commitment offers a further tool for correlated interaction in the face of the expanded cooperation enabled by commitment via shared activity. In other words, new forms of commitment coevolve with new forms of cooperation.[56]

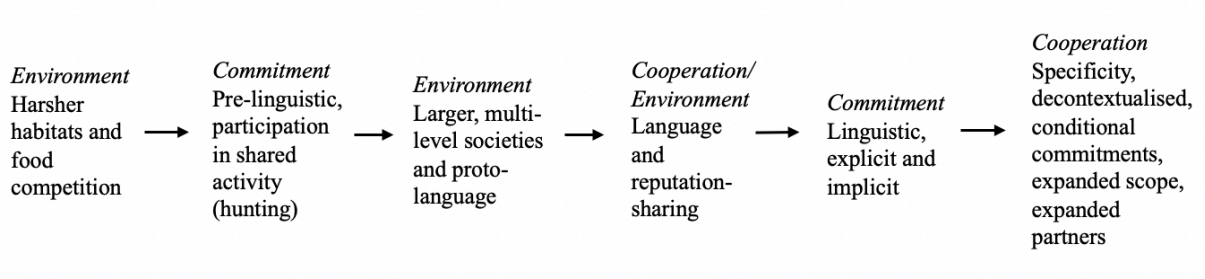

Figure 9: The coevolution of commitment and cooperation.

3.1 The emergence of language

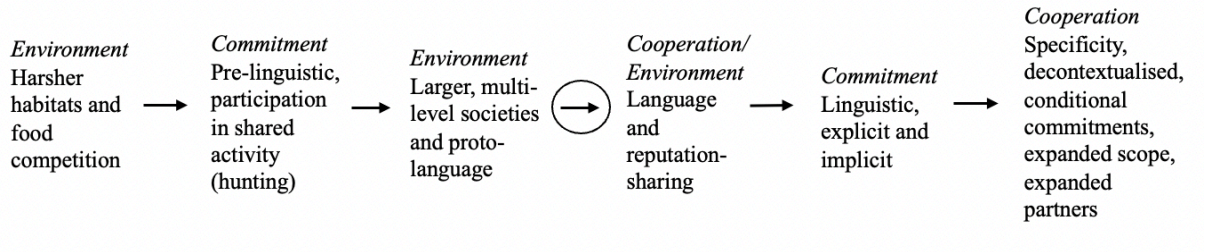

Figure 10: The emergence of language and reputation sharing.

Here, I detail how the requirement to share information in joint tasks with new interactive partners is a key contributing factor in the evolution of language. It is important to distinguish two types of claims I will be making about language evolution. One is about the cognitive *precursors* to linguistic communication and the other is about the *selection pressures* we faced which culminated

[56] Note that it does not matter for my thesis whether implicit commitment preceded or succeeded explicit promising. One might wonder how implicit commitment could be understood without a concept of explicit commitment already at play, but recall that commitment does not need to be intentional. All that is required for something to be a commitment is that it involves a pre-play signal which changes sender motivations and receiver expectations.

in language. When I claim that "*x* is a precursor to *y*", I mean that *y* would not have been *possible* without *x* capacity. When I claim that "*x* selects for *y*", I mean that *x* creates the environment in which *y* capacity becomes advantageous. Something can be a precursor without selecting for anything. For example, the ability to vocalise due to our particular anatomy is a precursor to natural language but this ability does not *select for* language – it is an environment where information sharing leads to differential reproductive success that selects for language.

In this section, I will (i) discuss one plausible account of the selection pressure for the evolution of our protolanguage reliant on joint activity; (ii) detail the cognitive precursors to abstract and decontextualised language; (iii) show how larger and more complex network structures created selection pressure for language evolution. Later in the chapter, I will suggest how coordination in more complex joint tasks creates selection pressure for the evolution of explicit promising, beginning with role announcement.

3.1.1. Joint tasks and gestural communication

On the account to be provided here, cooperative communication began with gestural communication aimed at coordinated action and later became the kind of natural language we are familiar with. I do not mean to suggest shared activity was the *only* important feature in the evolution of language. There are many differing accounts of the evolution of linguistic communication.[57] It is well beyond the scope of this chapter to survey them. Furthermore, it will not matter to my thesis exactly which selective forces played the more important role, since much of the coevolutionary story we are concerned with happens whether or not we view language as

[57] For a good overview, see Tallerman and Gibson (2011), as well as Pleyer and Hartmann (2019).

arising directly from commitment in the increased role for gestural communication in joint tasks. That is, it nonetheless holds that a new form of commitment – linguistic commitment – evolves out of the expanded cooperation that commitment via shared activity enabled. Since, as we have seen in the previous chapter, it is these activities which resulted in larger, multi-level groups and, I claim, these groups provide the selective environment for linguistic commitment to emerge to facilitate cooperation among new interactants (claim (iii) above).

However, the deeper causal relationship between commitment and language evolution (claim (i) above) will be well supported by any account that sees joint activity as a *driver* of language evolution, for example, those of Hurford (2007, 2011), Burling (2011) or Zuberbühler (2011). It will be less well supported by those accounts that do not see language evolution as dependent on coordination between individuals, for example, accounts which see grammatical structure as emerging in response to the individuals own cognitive needs to organise their language system in a user-friendly way (Goldin-Meadow 2003), or views which hold the primary purpose of language is not communicative but representational (Chomsky 1980; Bickerton 1990). However, communicative problem-solving likely did play a large role in language evolution (Sterelny & Planer 2021). One notable account of the evolution of language reliant on the role of joint activity is Tomasello's (2014). This is the account which I will use to illustrate the connection between joint activity and language, though it will not be the only account in support of the coevolution of commitment and cooperation.

Tomasello (2014) argues that where small-scale collaborative activities were required, natural selection favoured the development of gestural communication, pointing and pantomiming, which

were precursors to the development of conventional linguistic communication as we know it. He argues that, rather than the "requestive" pointing seen among great apes, the environment faced by early humans favoured the ability to develop and understand such gestures as directed at the achievement of joint goals. These gestures were not simply requestive, but could also be "informative" for another agent.[58] Tomasello (2014) argues that an agent understands why another is pointing at a particular object or pantomiming a particular task if both agents' attention is geared toward their collaborative activity. For example, for Sandy to understand why Betty points at the banana tree, she must understand that Betty is trying to inform her of something relevant to their current task of banana gathering.[59] Though I omit many of the details, Tomasello (2014) appeals to a number of studies on pre-linguistic infants to substantiate his thesis. To illustrate just one, Liebal and colleagues (2009) conducted a study in which 14-month-old and 18-month-old infants were tasked with cleaning up toys with the help of an adult and placing them in a basket. During this task, the adult had stopped and pointed to a toy, which the infants typically picked up and placed in the basket. However, when the infants were joined by a different adult with whom they had not engaged in a joint activity, the adult's pointing to a toy prompted the child to clean it up significantly less often. Instead, they offer the toy to the adult or point to it themselves. The proposed explanation is that their *shared common ground* is what allows the infant to understand

[58] Requestive and informative communication are delineated by Tomasello (2014) in terms of motive. Requestive communication asks for something while informative communication is intended to notify the receiver of something, ostensibly communication marked "for you".

[59] One might wonder why Sandy could not equally well take Betty's pointing to the tree to indicate that there are no predators, rather than that there are bananas. Tomasello (2014) argues this is because, in such contexts, the recipient infers that the communicator is communicating something *relevant* or *new* to her. If the agents have been engaged in a long search for bananas which has thus far been unsuccessful, Sandy is more likely to infer that Betty's pointing means that there are bananas in the tree. If the agents have recently been unable to acquire bananas because of the presence of a predator, her pointing might indicate that this is their opportunity to check the tree since there is predator around. The fact that we attend to what a communicator might see as relevant or new for us is evidenced in infant behaviour. Moll et al. (2006) conducted in study in which 18-month-old and 24-month-old infants played with an adult and a toy drum. If a new adult entered the room and pointed to the drum excitedly, the child assumed the adult was interested in the toy drum. However, if the adult with whom the child had just been playing pointed to the drum excitedly, the child investigated a specific part of the drum or looked around the room for a different object.

such gestures. If this hypothesis is true, the shared common ground created by joint activities such as hunting would have provided the selective environment for the emergence of gestural communication.

Note also that the importance of a shared common ground is not unique to a *gestural* protolanguage. This would equally well hold if the protolanguage is vocal. The key claim here is that joint activities are the drivers of our protolanguage, and this claim is supported by other accounts as well (Hurford 2007, 2011; Burling 2011, Zuberbühler 2011).[60] One might question why the joint activity is what enables Sandy to understand Betty. Why could Sandy not understand communicative gestures because she has seen others gather bananas? It is important to note, though, that if Sandy sees others pointing in such a way, they have done so because *they* are engaged in a joint task of banana-gathering and their communication is directed toward this activity. If she has seen and learns to imitate this gesture, we are on our way towards the conventionalisation of this signal. To argue she would use the conventional signal before it had arisen in joint activity is to jump the gun.

3.1.2 The cognitive precursors to language

The cognitive substrate that enables inferences from another's gestural communication to achievement of a collaborative goal is referred to as "joint intentionality" (Tomasello 2014). In particular, selection favoured the evolution of certain perspectival representational capacities,

[60] Indeed, Zuberbühler's (2011) account is vocal. Even if these accounts are false, and we ought to accept a more representation-first account of the evolution of language, this does not undermine the fact that a new form of commitment (linguistic commitment) evolved in response to a new form of cooperation (linguistic information-sharing), so much of the coevolutionary story is preserved.

socially recursive inferences and self-monitoring abilities in order to effectively engage in such communication. An individual takes on another's perspective in order to make their communicative act relevant to the other in the context of her goals and expectations. This also involves reflexively considering one's own communicative act in order to alter and better convey one's intentions. Furthermore, to communicate effectively, an agent must know that her partner intends her to know something, requiring at least one level of recursive inference. Such cognitive capacities are evidenced in studies with pre-linguistic infants.

A study by Liszkowski and colleagues (2009) with 1-year-old infants involved participants first repeatedly seeing an adult placing desired objects (toys) on top of one platform, while placing undesired objects (paper towels) on another platform, while another adult (verbally) chose the desired item. The items were then removed from sight and the demonstration repeated. In the test condition, the infants were placed in the receiver role and the item was no longer present on the platform. Many infants opted for pointing to the empty platform – to the location where, in their shared common ground, the infant and adult knew these desired items were generally found. Tomasello (2014) suggests that to perform this act, the infant is simulating the adult's process of comprehension – in effect, asking what abductive inference the adult will make if she points to the plate. The ability to view a situation from multiple perspectives and to adjust one's own expressions in light of this lays the groundwork for a transition to fully decontextualised linguistic communication.[61] Thus, the ecological changes that made collaboration important, notably in group hunting, would have created selective pressure for gestural forms of pre-linguistic

[61] For more on the cognitive abilities that make possible collaborative gestural communication, see Tomasello (2014).

communication (claim (i)). Not only this, but they would have also selected for the cognitive inferences we would need to understand and manipulate them (claim (ii)).

Yet fully abstract and decontextualised linguistic communication requires the development of certain cognitive capacities importantly continuous with the capacities developed in our protolanguage. In particular, in order to effectively communicate with in-group strangers – agents with whom we have never before interacted – it is necessary that agents are able to adopt an agent-neutral perspective or schematise certain situations in an agent-neutral way, using a culturally-constructed representational system, and to govern their thinking and communication by the normative standards of the group. This suite of capacities is referred to as "collective intentionality" (Tomasello 2014). The key claim here is that such cognitive capacities are *built upon* the joint intentionality of our protolanguage. Agent-neutral perspectives and group-mediated self-monitoring are natural extensions of second-personal perspective-taking and second-personal self-monitoring.

One might object that the literature on signalling games in the evolution of language provides an account of its emergence that does not rely upon any of the cognitive machinery that Tomasello relies upon (Lewis 1969; Skyrms 2002, 2010).[62] However, it is important to note that this literature primarily seeks to give a how-possibly account of the evolution of language, rather than a how-plausibly account. Though in principle it can be used to do so, the purpose of such models is usually to provide the minimal conditions under which a signalling system can evolve, rather than

[62] For a discussion on the adequacy of such modelling techniques in the evolution of language, see LaCroix (2020). The author argues that traditional signalling models are sufficient for explaining simple communication seen in nature, though the models would need to be extended and modified to account for the emergence of natural language.

to detail the most likely evolutionary trajectory we followed. One might argue that studies on infants employed by Tomasello do not provide good evidence of his how-plausibly account. This is because they may not track the cognitive capacities present in hominins in our ancestral environment. Indeed, the behaviour of infants might have been shaped by cultural learning from observation of adult behaviour, and adult behaviour may, in turn, have relied upon adaptations in response to the current environment – adaptations which would not have featured in our ancestral environment. However, the studies on pre-linguistic infants and their comparison with the behaviour of great apes represents the best evidence we have, and so should not be discounted.

I am not claiming that we ought to make inferences from facts about ontogeny to facts about phylogeny. Rather, from Tomasello's (2014) work, we can see that *certain sets* of cognitive abilities and cooperative abilities are tied together. That is, we are not making an inference from the abilities of children to our ancestors but rather from perspectival representational capacities, socially recursive inferences and self-monitoring abilities to the possibility of effective gestural communication, and from the capacity for agent-neutral thinking to the ability to communicate in a flexible and abstract way. That these two stages are distinct in our ontogeny is likely because of the complexity of the cognitive machinery involved at the different stages. It is this difference in complexity that lends weight to the idea that joint activities involving gestural communication arose phylogenetically prior to collective intentionality involving fully decontextualised language. With this in hand, we see how joint activities such as hunting likely enabled the cognitive capacities that underlie language.

The cognitive capacities that accompany Tomasello's account are intuitively crucial to any protolanguage that stems from ecological communication for joint tasks. Whether or not we took the precise steps Tomasello outlines, whether we believe "joint intentionality" captures a unified category of cognitive capacities and whether we believe vocal complexity preceded composite gestural signalling, will not matter. What matters is that shared activities such as group hunting selected for certain perspectival capacities, socially recursive inferences and self-monitoring abilities for communication that served as the cognitive prerequisites to abstract and decontextualised communication dependent on agent-neutral thinking, with communication governed by the normative standards of the group.

3.1.3 The selective environment for language evolution

We have seen how small-scale collaborative activities select for more informative communication in the form of gestures or pantomime. I now argue that small-scale collaborative activity also selects for decontextualised linguistic communication as a result of its contribution to growth in group size (claim (iii)). I suggest conventional linguistic communication developed in the context of increasing group size where it would be advantageous to have a group-wide common ground for coordinating action. Importantly, growth of the group sizes constitutes an environment in which we are potentially required to engage in cooperative interactions with *new partners on novel tasks*, making the pressure for a group-wide means of communication even greater.

To see why conventional linguistic communication is effective when interacting with new partners, we must first understand that the ability of early humans to communicate collaboratively through pointing and pantomime was useful when they shared a common ground of what was relevant to

the task at hand. Sandy understands Betty's pointing to the tree and pantomiming her climbing because she understands that, if Betty is directing her attention to the tree, it is because it contains bananas. Sandy knows this because their joint task of banana gathering is what creates the common ground of relevant communication. When the size of the group grows and one is forced to interact with unfamiliar partners or on novel tasks, this shared common ground may not be available. As a result, there is selective pressure to develop a decontextualised and more flexible form of communication to facilitate coordination with all members of the group.

Interpreting a gesture in a context where no prior communication has occurred requires a cultural common ground: things which members of the group know that others know, including an expectation about how people in the group communicate (Clark 1996). For example, Sandy moving her hand in a wave-like motion would only indicate the presence of a snake to a new interactant if this was conventionally understood as part of their group-wide signalling system (Tomasello 2014; Lewis 1969). These gestures can become increasingly arbitrary through stylisation and shortening, taken up in successive generations who acquire the communicative system culturally (Tomasello 2014). This process eventually results in fully decontextualised and arbitrary communication, or what we would naturally call a language. While the skills acquired in the group for social learning, such as imitation, might help to explain how pantomiming could come to be decontextualised and increasingly arbitrary, the details of how we arrive at such a signalling system is not of primary importance.[63]

[63] See Tomasello (2014) for a detailed exposition of how we may move from separate iconic gestures for acts such as opening a door and opening a jar, to an arbitrary sign for "open", to vocalisations which could represent abstract concepts such as justice, and finally to the schematisation of these abstract phrases into fully-fledged linguistic construction mediated by the social pressures of discourse.

What is important is that language *grew out of* the collaborative interaction underscored by the commitment mechanisms of the previous chapter. That is, it grew out of pre-linguistic shared activity based on commitment, establishing a coevolutionary relationship between forms of commitment and forms of cooperation. This is because shared activities likely selected for our protolanguage (claim (i)); they laid the cognitive groundwork for the development of arbitrary and decontextualised communication (claim (ii)); and they played a pivotal role in the development of larger, multi-level groups in which sophisticated communication became profitable (claim (iii)). As such, cooperative information-sharing via language coevolved with group hunting via commitment.

One worry is that it seems as though there would have already been language at play before the development of multi-level large groups. Indeed, how else would these groups communicate? However, first, if the account of Layton and colleagues (2012) from the previous chapter is correct, multi-level organisation could have developed as early as 800 kya, though estimates of language typically appear later than this (of course, dependent on what we mean by language). Powers and colleagues (2016), for example, estimate the arrival of language at 500 kya (though later estimates place it at 100 kya). Sterelny and Planer (2021) argue that language capacities were likely in place in the mid-Pleistocene ancestors of humans and Neanderthals, but that "modern" language did not appear until 200 kya.[64] Second, Layton and colleagues' account of the development of multi-level organisation sees this as arising from the growth and fission of larger groups. Though they do not discuss the drivers of larger groups, I have said something on this matter in Chapter 2. If this is the correct model of the development of multi-level organisation, then, it is likely that something akin

[64] See also Dediu and Levinson (2013).

to decontextualised communication began to arise before the fission of the village into smaller residential camps and facilitated communication across these camps, and this decontextualised communication was indeed driven by growth of the group size.[65] Yet the more that individuals interacted with members of other camps, the greater the need for this communication to become conventionalised in order to facilitate cooperation with new partners within one's wider social network. So language capacities likely coevolved with both the development of larger, multi-level groups and the maintenance of cooperative relations within such groups.

Where language is used to effectively engage in joint activity, it is itself a form of *cooperation*. Not only this, but conventional utterances (understood by all group members as part of the cultural common ground) enable much wider and more spontaneous forms of cooperative activity. The fitness advantages of sharing ecological information via language are clear. If information is passed on to collaborators in foraging efforts, or to offspring who are then required to spend less time learning before they can contribute to resources, it will be easier for members of the group to acquire resources, to avoid threats in their habitat and to survive (Sterelny 2012). Selection will therefore favour the persistence of traits that contribute to such sharing. Information sharing also proves advantageous with the development of a fission-fusion social structure, where separate forager bands explored different areas and returned to a central camp after their excursions, each possessing different information. Furthermore, it is possible language can also *incentivise* us to contribute in Public Goods games. For example, if linguistically codified norms of fair division exist (Smith 2010). So, not only does language prove a more effective mechanism for ensuring successful small-scale collaboration due to its specificity and ability to refer abstractly, but the

[65] Perhaps the fission also explains why many modern multi-level societies developed different dialects.

advent of language extends cooperation yet further – language allows us to cooperate in novel tasks with new partners.

3.2 Reputation sharing

As long as we are operating in an environment where we may have to interact with others with whom we have not interacted before, the ability to distinguish trustworthy cooperators from untrustworthy (or incompetent) defectors is advantageous. This creates selective pressure for language to be used in a different way, not only for the purpose of sharing ecological information, but also for sharing social information – the reputation of others as either cooperators or defectors. The selection pressure for a reputation sharing mechanism will be most salient in an environment where opportunities for commitment through shared activity such as hunting do not arise for all members, i.e. in large groups, so agents would benefit from some other means of determining a potential partner's reliability. If a practice of reputation sharing is in place, agents have a means of protecting themselves from exploitative partners. So, again, we see that commitment-based shared activity contributed to the formation of larger groups in which the sharing of information is required as a new form of cooperation. That is, communicative cooperation coevolves with commitment. The only difference here is that the information shared concerns social, rather than ecological, aspects of the environment.

The primary method by which reputation sharing occurs is through social "gossip" – the sharing of evaluative information concerning absent third parties (Foster 2004; Bergmann 1993). Dunbar (1996, 2004, 2011) (contra Tomasello's thesis) suggests this selective pressure for reputation

sharing is the reason that language evolved. He argues that as new ecological habitats were occupied by our primate ancestors, there arose new risks of attack from predators. As a result, group size increased to better protect its members from predators. This increase in group size entails an increase in the complexity of relationships that must be managed and comes with risks of feeding disruption and harassment from within the group. These developments favoured the evolution of displaced language to facilitate the sharing of information about those actors in one's network for whom direct monitoring of behaviour was not possible. This is known as the "gossip hypothesis". The thesis is congruous with the well-supported social complexity hypothesis for communication (Freeberg 2006; Kroodsma 1977; Pollard & Blumstein 2012; Wilkinson 2003; McComb & Semple 2005; Freeberg et al. 2011, 2012).

As noted earlier, it will not matter for my purposes whether Tomasello's or Dunbar's account of the evolution of language is correct. What matters is that earlier forms of collaborative activity create an environment in which a new form of cooperation is important for securing correlated interaction, whether this cooperation is in sharing ecological information in larger groups or sharing social information. If there is no direct link between our earlier small-scale collaborative activities and the development of language, our coevolutionary account will be one in which commitment via shared activity enabled the selective environment for language to evolve (as in claim (iii)), rather than an account about shared activities *selecting for* language evolution (as in claim (i)).[66]

[66] Ut is possible that Dunbar's gossip hypothesis may also rely upon the coevolution of commitment via shared activity and cooperation. In his case, the commitment would take the form of grooming and cooperation takes the form of the linguistic exchange of social information. See Dunbar (1996, 2004, 2011) for more details on the importance of grooming in the gossip hypothesis.

Linguistic reputation sharing confers an efficiency benefit since it is a means of pooling and exchanging information that does not require direct monitoring – it allows us to easily access information about other potential partners' characteristics and whether they are trustworthy without interacting with the agent. If Sandy hears that Danny often steals food after a collective hunt, Sandy knows not to cooperate with Danny, even if she has never interacted with Danny before. Indeed, in indirect reciprocity models in game theory, strategies which involve such reputation-based discrimination have been shown to result in evolutionarily stable cooperation (Alexander 1987; Pollock & Dugatkin 1992; Nowak & Sigmund 1998; Nowak & Sigmund 2005; Mohtashemi & Mui 2003; Panchanathan & Boyd 2003). The promotion of cooperation comes by way of gossip (Nakamaru & Kawata 2004).

Assessments of a potential partner's reliability would be furthered by developments in language capacities, such as the ability to convey information more granularly. For example, to say "Danny *always* steals food after the hunt" or "Danny only steals food on Thursdays" provides more granular information and thus furthers the potential for reliable correlated interaction. Cooperation is also furthered by decontextualised communication, since this allows us to communicate about spatially and temporally remote events. Such capacities enable access to information about our network in areas remote not only from our current interaction but also from our current environment. For example, Kenickie might hear that Danny always steals food after foraging even if Danny is not in his current hunting party. If Danny were to become part of his immediate environment, and seek to join his hunting party, he has reason to exclude him. Thus, the ways in which communication evolved over time to be more expressive and more decontextualised permits

more granular information sharing about the state of our social network via gossip, resulting in ever increasing capabilities for reliable cooperation.

However, if the sharing of reputational information via gossip is a means of securing cooperative interaction, should not the sharing of false information about others undermine its reliability? This is not necessarily so. There is psychological evidence that we have developed means of paying notice to the quantity, independence of sources, and motive of gossip statements in evaluating their veracity. Such mechanisms may play a role in stabilising the evolution of reputation-based cooperation even in the face of potential deceptive use. For example, agents may cross-check gossip statements against multiple sources; place more weight on the statements of direct observers than those of people with more degrees of separation; and discount information received from parties with personal interests (Hess & Hagen 2006; Sommerfeld et al. 2008; Smith 2003). Sommerfeld and colleagues (2008) find that whilst access to a single gossip statement does not lead to the same cooperative behaviour in experimental Trust games as direct observation of another's behaviour, access to multiple gossip statements does. This suggests that as the prosocial tendencies of the group incrementally increase, so, too, do their practices for protecting themselves from exploitation. Sterelny (2012) also emphasises the role of multiple information channels and network shape in maintaining the credibility of information sharing. Many-to-many networks are much safer than one-to-one information exchanges. Furthermore, to the extent that social information sharing is spontaneous and decoupled from immediate interaction, there is less opportunity to utilise this information for the exploitation of others.

Yet, if we already have a means of correlated interaction through reputation sharing, what good would commitment do? Reputation sharing does not render commitment redundant as an explanation of the evolution of cooperation, but rather bolsters it. Recall my aim is not to offer one canonical explanation of the evolution of cooperation but to suggest how commitment is one important piece of the puzzle, consistent with other accounts playing some role. Social information sharing and promising interact in important ways (regardless of which temporally preceded the other). We have seen that, in order for commitments to have fitness consequences, there must be potential for future beneficial interaction (either with the same agent or a different agent within one's network), or else there would be no cost to reneging. Promises might therefore already be made with familiar partners where future interaction is likely, and potentially some new interactants simply in virtue of being in a cultural group with a shared language. Importantly, though, access to an agent's reputation may bring new potential partners within scope of those with whom the agent is to have beneficial repeated interaction.

To elucidate: when gossip about a new partner is positive (or perhaps in the absence of negative gossip), this individual may be deemed trustworthy and thus a potential partner for mutually beneficial cooperative interaction. Gossip therefore extends the scope of partners with whom we make promises. Since Sandy hears that Betty is a cooperator, Betty becomes a more attractive candidate for repeated interaction, and it is possible to make commitments with her that change Sandy's fitness consequences for her future behaviour, since defection risks exclusion from current or future interaction. Indeed, other means of correlating interaction, such as group markings, may provide the environment ripe for effective use of commitment mechanisms in much the same way as gossip, by providing opportunities for future interaction. Furthermore, reputation sharing within

the group provides avenues for explicit promising to be profitable even where there will be no repeated interaction with the same individual. This is because a reputation for reneging on a commitment makes one vulnerable to exclusion from beneficial interaction with *other* members of the group, in congruence with indirect reciprocity models.[67]

3.3 The emergence of explicit linguistic commitment

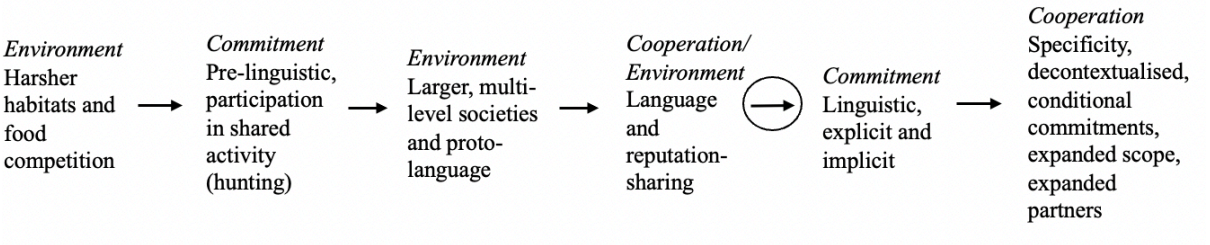

Figure 11: The emergence of linguistic commitment.

We have shown how our previous form of commitment created the selective environment for the evolution of language and reputation sharing. This new cooperative environment – one underpinned by ecological and social information sharing – makes possible the evolution of a new form of *commitment*, linguistic promising. As the size of the group grows, opportunities for shared activity with all members reduces. Under these conditions, there was selection pressure for a form of commitment that did not rely on prior interaction. In this section, I suggest explicit promising

[67] It is also important to note that credible commitments can be made even without one's reputation acting as the enforcement mechanism. More will be said on this in Chapter 5. In these cases, commitments may allow an agent with a bad reputation to recover their position in the social network. Thus, not only does the practice of reputation sharing among one's group form an environment in which correlated interaction among unfamiliar partners is possible, expanding the scope of potential partners, but commitment can also be a means of bolstering one's reputation and thereby increasing one's opportunities for cooperation.

arose out of the pressure to make clear the roles of interaction in complex shared activities. These complex tasks themselves arise because group hunting, among other factors, permitted an increase in innovation. Thus, we have a causal relationship between the requirements of coordination in shared activities and the evolution of a more sophisticated form of commitment.

As was discussed in the previous chapter, group hunting and the subsequent increase in the means of subsistence permitted an expansion in innovation. As a result of this and other factors, humans developed the more sophisticated tool-making and projectile weaponry of the Middle Stone Age. With the possibility for more complex tasks which might have involved several individuals, we face the problem of the division of labour. Here, some means of communicating one's intention in a joint task, such as making a sophisticated tool, would have made agents better off. Consider the anti-coordination game presented by jointly crafting a spear for hunting. The agents would do best to choose complementary actions (one crafts the spear shaft while the other crafts the spear head). The Nash equilibria here are (R1,C2) and (R2,C1) but there is no dominant strategy. As such, in the absence of a coordination device, each player on average yields a payoff of 2 rather than 3.

	C1 (Make spear head)	C2 (Make shaft)
R1 (Make spear head)	1,1	3,3
R2 (Make shaft)	3,3	1,1

Table 13: Anti-coordination spear-making game.

What is needed in a complicated joint activity of this sort is a *coordination device* or *symmetry breaker* (Skyrms 1996). That is, something which can be used to condition one's action on in order to move from an inefficient mixed equilibrium to an efficient equilibrium. Such symmetry breakers

might have initially been realised by way of natural salience. Suppose, for example, that Sandy was nearer the bananas in a joint foraging task while Betty was closer to the figs. It might have been natural for Sandy to gather bananas rather than Betty and this is what secured the mutually best outcome for them. However, roles in joint tasks were not always so clear. In tasks which were becoming increasingly artificial and complex, natural salience may no longer have sufficed to coordinate actions. As a result, there would arise selective pressure to create or use decontextualised and abstract communication to make explicit one's role in a joint activity.

If Sandy is required to make clear her obligations in the activity by communicating, for example, "I will craft spear heads while you craft shafts", she creates more grounded expectations of her behaviour in the cooperative interaction. What fitness advantage does this linguistic coordination device afford and how does this spread to fixation in a population? Those who make use of such a coordination device will have a fitness advantage over those who do not, since they will be able to coordinate on crafting weapons to hunt game that require a division of labour and divided expertise. To understand how this spreads to fixation in a population, it is useful to note that by the time we see advanced tool-making in the Middle Stone Age (and perhaps other complex tasks requiring coordination devices), unskilled individuals were learning by apprenticeship from experienced adults (Sterelny 2012). That is, experienced craftsmen led the instruction of the unskilled both through scaffolding their environment for easier learning and by direct teaching. The full details of the apprentice-learning model are not important for my account. What is important is that if younger individuals learnt from experienced members of the group, they would witness language being used as a coordination device in joint activities that require a division of

labour. If this is so, successful strategies will be copied and proliferate over generations, such that making one's role clear becomes common practice.

Why could we not simply rely on a convention as a symmetry breaker? That is, perhaps an accidental regularity in behaviour is all that is needed and, once established in the population, it would persist. Sandy would always be the spear head crafter while Betty always makes shafts. I argue conventions are insufficient on two grounds. First, role ambiguity becomes a problem as the cooperative tasks we come to be engaged in increase in complexity. Conventions about past roles will no longer suffice in novel cooperative enterprises and this is particularly so as the scope of potential interactive opportunities increases. This increase in complexity is a result of successful coordination in more rudimentary tasks as these more rudimentary tasks provide us with the means of subsistence to begin exploring and innovating. Consider, for example, the amount of activities possible on a farm once we have developed agriculture. Given that there are tens of tasks to do, clear communication becomes an important way of coordinating action. Second, with an increase in complexity comes an increase in the need for plasticity in response. Linguistic communication about role expectations allows us to respond flexibly in the face of contingencies we have not before encountered – for example, coordinating a flood response on the farm – in a way that mere conventions could not have. Thus, such role announcement in an anti-coordination game has clear fitness advantages compared to reliance on conventions.

Recall that what makes an utterance a promise is that one would face a cost for acting in contradiction to one's utterance. Note that if Betty's expectations of Sandy's future behaviour change, Sandy's cost of defection is raised, since if she does not follow through on her word, Betty

views her as a less desirable partner for future interaction. Thus, we have an initial form of promising in verbally making clear the division of responsibilities in shared activity. Fully fledged promises, however, will be detached from shared activity – they will be able to refer to abstract activities which we have not yet engaged in. I have nonetheless offered an explanation of how our very first explicit promises may have arisen and the connection this has to our previous forms of cooperation.

Of course, there are intermediate stages between pre-linguistic commitment and commitment dependent on fully decontextualised, abstract language. Our promises will develop as our communicative capacities develop in scope and complexity. Simple gestures may communicate one's intention in sufficient detail for commitments to be made in cases of joint activities. Composite gestures will allow additional complexity in communicating one's intention – not only may the sender of the signal communicate what she intends to do, but also by what means, through stringing together iconic gestures (Sterelny & Planer 2021). For example, one can signal a commitment not only by showing up to the hunting excursion but also by pantomiming one's intended role in the hunt or pointing to where one intends to go. Such composite signals would aid in receiver expectations of the sender's cooperation, securing mutual benefit by way of a more sophisticated commitment. The effectiveness of commitment will increase as our protolanguage develops in flexibility and precision. That is, our language capacities likely coevolved with the explicitness and efficacy of commitment. In this chapter, we will focus on the operation of commitment once we have fully decontextualised and abstract communication at play since this is where we see the largest change in our cooperative capacities.

3.4 Explicit linguistic commitment

What makes promising a commitment is that, if the promise-maker subsequently breaks her promise, she may be deemed untrustworthy and excluded from future beneficial interaction. As with commitment via shared activity, explicit promising serves the same two functions of introducing a fitness cost to reneging for the sender and facilitating the identification of cooperative partners for the receiver. Here, the commitment does not itself generate opportunities for future beneficial interaction since it does not (necessarily) involve a shared bonding experience. Rather, opportunities for beneficial interaction – with the same agent and with others – are available in virtue of an agent's reputation. Agents can use this to identify reliable partners for interaction. Not only this, but the ability to commit and follow through on one's commitment may act as a means by which one's reputation can be altered since news of one's actions will spread. In this section, I detail how explicit linguistic commitment operates and provide evidence for this conceptualisation of its operation.

How do we fit explicit, linguistic promising into our framework from Chapter 1? First, recall the definition of a commitment: *a commitment is a pre-play signal in a strategic interaction taken at time t, that increases the sender's relative payoff for carrying through option X at time t+n, and increases the receiver's probability of the sender carrying through option X.* In the case of explicit promising, we are not only confined to Stag Hunt games but any game in which there is mutual benefit to be gained from cooperation. As such, promises to cooperate can also occur in Prisoner's Dilemma contexts or Trust games. The move that the agent takes at time *t* is the explicit promise, which will usually take the form of an utterance such as "I promise to help" (the cooperative

option).[68] Option *X* – the option that the agent ensures she carries through with greater probability – is the cooperative action. This is Stag in the Stag Hunt, Cooperate in the Prisoner's Dilemma, or Split in the Trust Game. Time *t*+*n* indicates when the game is played. The pre-play signalling alters the payoffs of downstream choices, incentivising acting in line with one's promise. The incentive to cooperate depends on an increased cost of defection as a result of the pre-play signal. This cost of defection takes the form of either potential exclusion by the partner one has defected on in repeated interaction or exclusion from others as a consequence of a tarnished reputation.

The one-shot extensive form game representation in Chapter 1 codifies the change in subjective payoffs as a result of commitment signalling. That is, it captures the proximate mechanisms associated with ostracism or a negative reputation. The ultimate causes are captured in the evolutionary game. Here, the payoff change is seen in the agent's reduced opportunities future interaction with others. All of the parameters in the simplified evolutionary model presented in Chapter 1 remain the same but we may add additional nuance to their interpretation. In particular, *c* now captures the cost of exclusion not only from interaction with one's current partner but also exclusion from others on the basis of one's reputation – with a bad reputation, *c* increases, since *c* captures the cost of foregone opportunities for interaction. As noted in Chapter 1, in an evolutionary model, this would not be captured in one term but would rather play out in the dynamics. That is, agents who choose the strategy *SR1 & R2* would be less likely to be chosen to play the game in subsequent rounds. We can see how explicit promising against the backdrop of social gossip increases the stakes for following through on one's commitment.

[68] Of course, we now understand locutions such as "I will help" without the preceding "I promise" phrase as commitments, too. How we came to understand these phrases as meaning the same thing is not of importance.

In the previous chapter, we provided evidence that showing up to a hunting excursion is an instance of commitment by showing that this generates future opportunities for beneficial interaction and that reneging on one's commitment is punished, increasing the cost of defection. For the case of explicit promising, I hold that, once larger cultural groups are established and reputation sharing is in place, it is not necessary that the commitment itself generates future opportunities for beneficial interaction. This was previously important as reputational information was unavailable to an agent. However, now, opportunities for future beneficial interaction exist in virtue of the interconnected practices of the group and in virtue of shared knowledge of potential partners' trustworthiness. That is, with interconnected groups and reputation, we do not need social bonding or past observation of play to make a commitment function, since reneging introduces a cost to reneging via tarnishing one's reputation among one's wider network of potential interactants – not only by fracturing a particular trusting relationship. We must still show that promise-breaking is punished.

The disposition to punish betrayal is evidenced in experiments, especially in the presence of broken promises (Boles et al. 2000; Koehler & Gershoff 2003). Koehler and Gershoff (2003) find that people react more strongly, both in terms of punishment enacted and negative emotions felt, to acts of betrayal than to identically bad acts that do not violate a duty or promise to protect. In the studies, participants were presented with descriptions of five different crimes and were asked to recommend jail time punishments for each perpetrator of the crime. In cases where the perpetrator's profession indicated that they ought to protect against the crime, their action was punished more severely. Furthermore, participant's written explanations pointed to the *broken promise* as the largest effect on recommended jail time. This is in contrast to increased

recommended jail time stemming from the perpetrator's ability to avoid detection when they occupy such a role or stemming from exploitation of thier special access. This shows that we do in fact punish others for broken promises, increasing the cost of defection after commitment signalling. These negative attitudes toward promise-breaking were likely the result of adaptations to limit agent's susceptibility to deception, discussed in detail in the next section.[69]

Furthermore, unfair decisions in Ultimatum Games are met with higher punishment rates when preceded by deceptive messages than when they are not (Brandts & Charness 2003). This means the fitness cost of defection is greater in the presence of a prior commitment signal than in its absence. That is, commitment raises the cost of defection. This accords well with the simplified model presented in Chapter 1, suggesting that $EV(SR1 \ \& \ R2) < EV(\sim S \ \& \ R2)$ in virtue of c, so sending the signal *SR1* in the pre-play stage of the game incentivises also choosing *R1* in the second stage, since those who do so avoid the penalty, c, if their defection is detected and so fare better in the evolutionary dynamics. If this is the case, one might wonder why the defector would signal at all. As noted in Chapter 1, the strategy *SR1 & R1* increases $P_C(R1)$, incentivising *C1*. If Row is a defector, this would ensure them a higher payoff in the Prisoner's Dilemma – the "temptation" payoff rather than the "punishment" payoff. This is similarly so in the Trust game, since one receives a proportion of the trustor's endowment.[70]

[69] While the negative feelings we have toward those who break promises which directly affect our welfare is clearly adaptive, the feelings we have toward promise-breaking between *third* parties that does not affect us is not obviously so. Nonetheless, it is not uncommon that cognitive adaptations extend to domains other than those of their initial use.

[70] Note that a Hare-hunter signalling Stag would not secure them a higher payoff in the Stag Hunt, since $z < y$. However, in reality, group hunting is best modelled by a n-player Stag Hunt or Public Goods game in which there is something to be gained by signalling cooperation if one is a defector – one gains the opportunity for free-riding. This is discussed in more detail in Chapter 2.

I suggested that what incentivised promise-keeping at the proximate level in the case of hunting was social bonds. With explicit promising, other proximate mechanisms become relevant, and these may operate in the absence of social bonds. For example, promise-keeping may be incentivised as a result of concern for one's reputation and avoidance of ostracism, for which there is ample psychological evidence. Emler (1990) presents experimental evidence of active management of one's reputation. Participants in the study were asked to imagine themselves involved in various events, either good or bad (for example, winning a scholarship or being in a motor accident) with varying levels of responsibility for the event. They were then asked to generate a list of people they knew in different categories of social relationships and to indicate how much effort they would undergo to tell these people about the aforementioned events. Subjects made more effort to communicate information likely to be to their credit, and the events they shared depended on their closeness to the person with whom they shared. Participants were more likely to share potentially damaging information with family and close friends. Of particular interest is that when the situation was manipulated such that incidents could be witnessed and reported to others, it was found that participants were much more likely to reach friends with "their own version of events" (Emler 1990: 184) and this was particularly true the denser the network of connections. There is therefore evidence that people engage in proactive reputation management out of concern for their appearance to others.

Gruter and Masters (1986) consider the adaptive role of ostracism, arguing that groups which ostracised burdensome or deviant members became more cohesive, their members enjoying the benefits of greater security. Williams (2007) notes this selects for an ostracism-detection mechanism. Organisms that were especially good at anticipating ostracism were better placed to

mitigate the loss of group membership, and this may favour a bias toward false alarms rather than misses, meaning we are highly attuned to potential ostracism. In experiments by Stroud and colleagues (2000), researchers found that excluded participants experience significant increases in blood pressure and cortisol levels as well as greater self-reported levels of tension. In experiments by Eisenberger and colleagues (2003), ostracism was associated with increased activation of the dorsal anterior cingulate cortex, a region of the brain that shows activation during exposure to physical pain, and was highly correlated with self-reported distress. Chen and colleagues (2008) also find that individuals can relive and reexperience social pain both more intensely and more easily than physical pain.

Finally, Haley and Fessler (2005) present a series of experiments that test for reputational concern in the motivation of cooperative behaviour in Dictator Games. They manipulated whether the desktop background displayed two eye-like shapes or the laboratory's logo and found that participants gave more money in the former case than in the control condition, suggesting the eyes acted as a subconsciously understood observation. The same findings were also replicated with noises or silence – participants gave more money with the presence of background noise. These studies show us that ostracism has a significant impact on a person's happiness and that avoidance of such an outcome may well be a proximate mechanism incentivising an agent to follow through on their commitments.

Recall these proximate mechanisms are those which impose an intrinsic disutility to defection for the agent who has signalled her intended cooperation, motivating cooperative behaviour. The additional social sanction in response to defection is what makes commitment behaviour fitness-

enhancing from the perspective of ultimate causation, as well as what secures its credibility, but need not be what directly motivates the agent to act in line with her signalled intent. One might worry that such proximate mechanisms will not have been available to our ancestors – that fear of ostracism is a late-emerging phenomena. However, all that this requires is a concern for what others think of us. Given that other great apes are aware of their ranking in dominance hierarchies, they certainly have an awareness of what others think of them. Not only this, but there is evidence of apology and amendment in reconciliatory practices (Ostner 2018; Arnold & Aureli 2007). For example, victorious chimpanzees in dominance battles will often work quickly to reestablish a relationship with the defeated male in order to secure a tenuous position (by offering their rear-end and engaging in grooming) (De Waal & Van Roosmalen 1979). If the capacity to recognise one's relative ranking and to make efforts to sustain one's position is present in chimpanzees, this is likely also present in our hominin ancestors. This suggests that concern for one's reputation is an available proximate mechanism for securing commitment when we evolved language.

So promising functions in a similar manner to commitment based on shared activity in that it signals cooperation to others and it is fitness-enhancing for the agent to act in line with her signalled intent to cooperate since defection may entail exclusion. Access to reputation also expands the scope of partners with whom commitments can be made, since it can render a partner more attractive for future cooperative interaction. Since gossip likewise affords others access to one's own reputation, commitment can also be used as a form of reputation management. To undertake and follow through on a commitment establishes one as a cooperator in the network and increases one's opportunity for beneficial gains. Defection now also entails news of the agent's defection reaching other potential partners – not only the agent who has been exploited –

compounding its fitness costs and further favouring cooperative behaviour in the presence of a commitment.

3.5 Implicit linguistic commitment and its emergence

We have seen that our previous practices of commitment created the environment, and selective pressure, for the evolution of language. Language was useful in the sharing of ecological or social information. Once we were able to share social information about others, appraisals of their actions could become meaningful for our future behaviour. In this section, I elucidate how an implicit linguistic commitment works and discuss how we might have transitioned from our ordinary information-sharing practices to the making of implicit commitments on the basis of gossip. This relies on the emergence of shared norms of behaviour. It is well beyond the account to offer an explanation or periodisation of the emergence of normative thought or cultural norms. So, for now, let us assume that we have this apparatus at play and examine the consequences that cultural norms had for our forms of commitment.

Before continuing, it is worth clarifying some terms. I will define "norms" only to the extent that we need them to talk meaningfully of new forms of commitment and cooperation. According to Schlingloff and Moore, "a norm is a rule that agents feel, in some sense, obliged to follow" (2018: 381). Indeed, in the traditional philosophical literature, a nomic capacity refers to the capacity to "follow rules" and is thought to be a quintessential human capacity (Hayek 1982: 11; Searle 2003:

200).[71] Of course, to possess a concept of prescribed rules does not entail that we always adhere to them. It means that we are able to take them into account when making our decisions, even if our decision is to break the rules. As such, a nomic capacity is best understood as the ability to act *in light of* rules (Lorini 2018).[72] Such rules are united by their prescriptive, permissive or prohibitive qualities, rather than simply being descriptive of agents' behaviours.

Under this definition of a norm, it is relatively uncontroversial that many early tribal societies possess cultural norms, for example, in dress, ritual or kinship relations. The norms are local, in that they do not apply to members outside of the group, and they are often self-imposed. Since these norms constitute part of the identity of the group, they are often also motivationally powerful. A person from a particular tribe may take pride in their adherence to the ritual practices of that tribe since it constitutes part of their identity. Not only this, but these are rules which agents feel, in some sense, obliged to follow. In this context, social information-sharing will have important consequences. Agents can begin to use social information not only to learn about the target of the gossip, but to make inferences about the normative attitudes of the person who shares such information.

I show how gossip concerning a third party's adherence (or lack thereof) to a norm can be taken as an implicit commitment to behave in a similar/dissimilar manner to the target of the gossip. Gossip statements will operate as commitments as long as the receiver believes the sender to share

[71] There is considerable controversy over whether animals exhibit social norms. Whether they do so does not matter for this thesis, since we are concerned with the evolution of norms within the hominin lineage. For overviews on this topic, see Liorini (2018), Schlingloff and Moore (2018).

[72] Bicchieri (2006) defines a norm as "a rule of behaviour such that individuals prefer to conform to it on the condition that (a) most people in their reference network conform to it (empirical expectation), and (b) that most people in their reference network believe they ought to conform to it (normative expectation)." This definition will also work perfectly well for my purposes.

the same cultural norms and, where relevant, occupy the same social role as the agent about whom they are gossiping. If Sandy says to Betty, "it is terrible that Danny did not attend the fortnightly ritual dance" and Betty believes that Sandy is a person who shares Danny's cultural background and occupies a social role that involves her attending the fortnightly ritual dance, too, Sandy's utterance to Betty can be taken as an implicit commitment not to behave in a similar manner. The gossip signals her attitude toward the behaviour of the gossiped party. Notice that it is not enough that Betty understands Sandy to be part of the same group if attendance of the fortnightly ritual dance is a role-dependent activity. In this case, she must also believe that Sandy's social role is relevantly similar to Danny. Suppose, for example, that only men ought to attend the fortnightly ritual dance. If this is so, Sandy's utterance does not imply anything about her future behaviour since the norm does not apply to her.

To elucidate with a more modern example, suppose Sandy said to Betty, "I believe that it is terrible Danny ate bacon even though he is Jewish." Suppose Sandy takes it to be the case that *only Jewish people* should not eat bacon and would not so criticise a non-Jewish person for behaving in this way. If this is so, the only feature of this gossip statement that makes Sandy liable to future exclusion is if the receiver identifies her as part of the same cultural group – as Jewish. Whether or not she truly is part of the same cultural group does not matter, since Betty will take her to be a hypocrite insofar as she believes Sandy is Jewish. Suppose that Betty believed that Sandy was not Jewish and believed that Sandy would not so criticise a non-Jewish person – Sandy's gossip statement would then imply nothing about her future behaviour since she does not take this norm to apply to non-Jewish people. In this way, group-specific norms can imply meaningful commitments on the part of the gossipers as long as their group identification and social role is

believed to be in concord with the gossiped party. We expect receiver judgments about this to be made on the basis of some evidence rather than arbitrarily, so it is likely that, more often than not, these judgments are correct. Note that, since these judgments are not entirely up to Sandy, implicit commitments need not be intended.

With the relevant inferences concerning Sandy's shared membership in the cultural group and that she occupies the same social role, the attitude Sandy expressed is understood to apply to her as well, such that subsequently acting in contradiction to the signalled attitude would render her vulnerable to exclusion from future interaction on the basis of hypocrisy. That is, if she acts in contradiction to her signalled attitude, she reveals herself to be a hypocrite and thereby becomes a less desirable partner for future interaction. So, in conjunction with the adaptations we have developed for the detection and punishment of false signalling, Sandy's utterance operates as an implicit commitment. It introduces a fitness cost to reneging from one's signalled intent. As with explicit linguistic promising, opportunities for future interaction exist with familiar partners on the basis of valuable relationships as well as unfamiliar partners rendered more attractive on the basis of their reputation. Exclusion is therefore detrimental and one is incentivised to follow through on one's signal.

In the language of our definition from Chapter 1, *a commitment is a pre-play signal in a strategic interaction taken at time t that increases the sender's relative payoff for carrying through option X at time t+n, and increases the receiver's probability of the sender carrying through option X.* Here, the move that the agent takes at time *t* is the gossip concerning a third party. The option *X* that the agent ensures she carries through with greater probability is either: the action which she

has praised another for, or avoidance of the action which she has criticised. Time $t+n$ is when the opportunity arises for her to engage in or avoid the action she has gossiped about. Gossip therefore acts as a pre-play signal which alters the payoffs of downstream choices, incentivising acting in line with one's implicit commitment, option *X*. The extensive form representation in Chapter 1 captures the subjective payoff change as a result of the proximate mechanisms of commitment. The evolutionary implications are captured in the simplified evolutionary model presented in Chapter 1.

Of course, in order for such gossip statements to count as implicit commitments – especially when the behaviour in question might not be observed in the immediate future – the receiver needs to remember the information contained in the signal at some later time $t+n$. This is because, at time *t*, Sandy might not imminently face a ritual dance or the possibility of eating bacon. Yet, we have psychological evidence that gossip is salient to us and retained at a higher rate than other information. In a study conducted by Rogers (2017), participants were presented with nine paragraphs of written information, including scientific information and celebrity gossip. Participants rated the information on a scale of 1-10 based on personal relevance and how interesting they found it. After a week delay, participants were presented with a multiple-choice memory test about the information. It was found that, while scientific information was the most relevant and there was no significant difference in interest, celebrity gossip was remembered at higher rates than scientific information. As such, gossip statements may usefully be seen as commitments across time. The agents with whom we typically interact are not celebrities, but the study suggests that the important feature here is familiarity. Of course, implicit commitments will

be most useful in ensuring correlated interaction when the subject of the gossip concerns an activity the partners are imminently engaged in, or when the time between t and $t+n$ is short.

It is likely gossip about cultural norm violation first emerged as a correlation device before we understood it as a commitment (in the sense of changing sender payoffs as well as receiver expectations). In passing a descriptive judgment of another, for example, "Danny did not attend the fortnightly ritual dance", Betty is able to identify Danny as either a preferred or unpreferred partner. This is useful for Sandy since in trading information, she is more likely to receive some information back about the state of her social network and therefore to choose partners more wisely. However, notice that when Sandy's utterance is not only descriptive but expresses an *evaluative* attitude toward Danny – when she says "it is terrible that Danny did not attend the fortnightly ritual dance" – Sandy may reveal information about herself that allows Betty to correlate her interaction with not only Danny, but Sandy, too.

We can explain the capacity of this utterance to achieve correlated interaction by appeal to the value of information-carrying signals in Skyrms' (2002) sense. A signal carries information about an agent if the probability of the agent being in a certain category conditional on the signal is different from the probability that she is in the category based only on the population distribution. Suppose this kind of information is present at the beginning of an evolutionary process: for example, it is more likely that a person is a cooperator given that she gossips negatively about Danny's defection. This fact might even be pure accident. Skyrms (2002) shows that populations which employ strategies that exploit this information will grow faster by adaptive processes than

populations that do not. This is because these agents cooperate only with trustworthy partners on the basis of this signal.

In this way, agents who were able to utilise the information present in gossip would have received more cooperative benefits than agents who were not. They not only shared and received social information about others but simultaneously used gossip about cultural norm violations to discriminate among potential partners. They would have thereby better avoided exploitation by defectors and secured interaction with cooperators. The process of exploiting information explains how such gossip could be used to correlate interaction. Note, also, that this explanation does not require that all people in the community use such norms as a means to securing effective partner choice. All that is necessary for this tendency to spread is that some do, and those individuals fare better by correlating interaction.

Such correlation devices just are commitments as long as reneging is punished. In these circumstances, gossip about cultural norm violations not only change a receiver's expectations about the sender's likelihood of cooperation, but also the payoffs of the sender in virtue of potential exclusion from future mutually beneficial cooperative gains. That is, once the practice of inferring the gossiper's attitudes from their gossip becomes widespread, general expectations may form about what can be reliably inferred from various kinds of normative criticism and failure to adhere to the inferred expectations can be seen to be in violation of an implicit standard set for one's own future behaviour. If an agent is seen as a less desirable partner on the basis of this violation, there is a cost to reneging, characteristic of a commitment. The cost of defection depends on future opportunities for beneficial interaction which will be compromised if the agent is a less desirable

partner. Such opportunities will be rife where we have shared social networks with reputation sharing at play.[73]

Echoing our discussion from the previous chapters, we see the evolution of a signal from a cue. Once it is understood that the signals which initially were designed to transmit information concerning third parties are also cues to our own future behaviours, agents can begin to use these gossip signals to manipulate the behaviour of the receiver, and to intentionally signal their future behaviour. If gossiping about the behaviours of a third party increases the likelihood of cooperative behaviour on the part of her audience, the sender's disposition to gossip is reinforced. The sender could then manipulate the behaviour of the receiver by choosing to convey information about their intended cooperation via gossiping about others. Since this increases the receiver's likelihood of cooperating with them in turn, it secures mutual benefit. In this way, we may see how evaluative gossip comes to constitute a commitment.

Indeed, in an environment where reputation sharing is a widespread form of correlating interaction, selection favours using this mechanism not only to glean information about the gossiped but also the gossiping agent. That is, it favours a new form of commitment to determine the trustworthiness

[73] While this might initially appear to render a commitment no different to a correlation device, in Chapter 1, we saw that many explanations of the evolution of cooperation espouse a way of correlating interaction. The difference is in the means by which we achieve correlated interaction. In the case of commitment, the correlated interaction is achieved by means of a pre-play signal which changes the sender's relative payoffs. In the case of reputationally-backed commitments, it is the change in the receiver's expectation which causes the change in sender payoffs, since the receiver is prepared to exclude the sender on the basis of hypocrisy. As such, signalling cooperation and subsequently defecting incurs a fitness cost (notice that it is not defection alone that results in the change of sender payoffs, but specifically the violation of the signal to cooperate). Here, commitment resembles and combination of pre-play signalling and indirect reciprocity. However, this will not be true of commitments more generally. In cases of financially or physically secured commitments, the change in sender's relative payoff is undertaken exogenously then signalled to the receiver, and this is what changes the receiver's expectations. The change in the sender's payoffs here is not dependent on the change in receiver expectations since the commitment is enforced by a third party.

of potential partners and to advertise one's trustworthiness to others. It is against this background that cultural norms can make possible implicit commitment on the basis of gossip. After the advent of cultural norms, gossip which had previously been used for reputation sharing purposes is now also understood as an implicit commitment to act in similar/dissimilar manner to the praised/criticised party.

3.6 The advantages of linguistic commitment

Though linguistic commitments as I have outlined them operate in the same basic way as commitment based on shared activity (via increased sender cost of defection and changes in receiver expectations), language may be a more *effective* means of signalling commitment and this increased effectiveness will explain its emergence where other forms of commitment already exist. This is particularly so where opportunities for shared activity may not be available in a larger group. Below, I illustrate the fitness advantages of linguistic commitment, which apply equally well to explicit and implicit promising.

First, language offers an efficiency benefit in communication over pre-linguistic commitments even among familiar partners and in the same shared activities. One benefit derives from its generality and one from its capacity for specificity. In being a decontextualised form of communication, the same linguistic utterance can serve to facilitate cooperation in multiple contexts. For example, the ability to say "I will help you" signals cooperation in both alloparental care and hunting situations. In contrast, some pre-linguistic signals of cooperation, for example, the staccato grunting of stump-tailed macaque alloparents or "hugging" of New World spider

monkey alloparents will not be a sufficient signal of cooperation in a hunting scenario (Henzi & Barrett 2002; Hrdy 2009; Bauers 1993). Therefore, language can extend the range of matters on which we can undertake commitments without need to develop a new commitment vocabulary, since language makes use of general and decontextualised signals of cooperation.

As our vocabulary and syntax develop, language also allows us to communicate commitments with more specificity. Communicating a commitment as specific as "I promise to collect figs while you forage for bananas" allows the receiver to form more grounded expectations of cooperation concerning actions she may not directly monitor. Likewise with implicit commitments such as "it is terrible that Danny did not collect figs while you were out foraging for bananas". Linguistic commitments are thus more effective in changing the receiver's expectations than commitments based on shared activity, where all that is revealed about the partner is that they are trustworthy in some general sense – that they would not exploit the actor in times where coordination is required, not that they would follow through on a particular action which they have specifically communicated. The aforementioned communicative capacities expand the range of matters on which we can commit.

Furthermore, conditional commitments are now possible. These are modelled by a Trust game. As a reminder, in the Trust game, players have an initial endowment and the trustor has the option of sharing a proportion of her endowment with the trustee. The sum sent to the trustee is multiplied. The trustee then has the option to share some of the multiplied sum in turn. It seems this fits the structure of many ordinary cooperative interactions: if trusting another agent is risky since she may not repay us in kind, we would need from her a credible, conditional promise to cooperate if we

do so first. Linguistic promising allows us to make promises of this form. In a multi-stage interaction, it allows the second player, the trustee, to say: "if you share *x* proportion of your endowment with me, I will share *y* with you in turn". This will benefit both players since the endowment is multiplied, incentivising the trustor to share *x.* Promises of this form were not possible with our earlier form of commitment based on shared activity. This means the range of cooperative enterprises now includes those interactions in which partners have an exit option that does not hinder their fitness, but also an option for a mutually beneficial but risky outcome, and where they have the option to move first conditional on the commitment of a second player. Many real-life scenarios take this form: the agent can choose to interact with another or not, but interaction is only worthwhile if the agent can trust the other. Thus, language widens the scope of potential cooperative interactions in which commitment can be meaningfully employed.

Conditional commitments will also enable us to strengthen our promises by specifying the consequences of noncompliance. With language, Sandy can now communicate to Betty "I promise to collect figs and if I fail to do so, you are entitled to take my bananas." In this way, language allows us to specify an additional means of enforcement which can make one's commitment more credible. If such commitments are made publicly, Sandy's reputation will suffer not only if she reneges on the commitment to gather figs, but also if she does not allow Betty to take her bananas. Betty therefore has reason to believe that Sandy will follow through on her commitment or compensate her appropriately. Through use of these qualifiers, it is possible that even if Betty did not believe Sandy's previous fig-collecting promise, it can be sufficiently strengthened such that she will. Why would this promise be more credible? If the agent herself can specify the contingent costs of her defection, and does not abide by this specification, others can not only point to her

failure to carry out the promise but also her failure to abide by the stated penalty, tarnishing her reputation further if she reneges. Since her cost of defection has increased, the agent's promise is more credible. Strengthening promises in such a way can create new opportunities for mutually beneficial interaction, since promises which were previously not believed may become so. This extends the scope of cooperative endeavours we can engage in. Simultaneously, it creates new opportunities for testing a potential partner's reliability which better enables us to identify cooperative partners for interaction, particularly in the context of growing group populations where shared activity and social bond creation may not always be possible.

Implicit commitments can also be conditional and can specify the penalty for reneging. We may gossip about someone who has engaged in transgressive behaviour whilst also specifying the conditions under which we would consider it acceptable. For example, in the context of food-sharing norms, Sandy might tell Betty, "it is terrible that Danny stole all the figs when Kenickie gave him those bananas." This communicates to Betty that Sandy is committed to not stealing figs, *if* she receives bananas. Sandy might also say "it is terrible that Danny stole all the figs – someone who does that should not get any bananas the next day." Here, she specifies the further contingent costs she expects to pay upon defection if she acts in a similar manner, in addition to potential exclusion (she expects to pay these as long as the receiver believes she is of the same cultural group and occupies the same social role as Danny).

Additionally, the ability to communicate about spatially and temporally remote events offers advantages for cooperation. The capacity to make promises on issues outside of our current cooperative endeavour creates even more opportunities for mutually beneficial interaction. Sandy

can now communicate to Betty: "I promise to collect figs on our foraging excursion next week" or "it is terrible that Danny did not collect figs for Kenickie when he told him he would do so a week ago". Since the receiver is only able to verify whether the sender follows through on her commitment at a much later date, it might be thought that these kinds of commitments could not be credible since it is unclear whether defection could be observed. However, with language on the scene, this can be verified through gossip (analogously for promises concerning spatially remote events). Not only this, but an agent's reputation can render their commitment credible where direct verification is unavailable.[74] So, provided they are deemed trustworthy on the basis of reputation or prior engagement, commitments regarding spatially or temporally remote events will extend our potential cooperative enterprises further.

Finally, language enables us to determine the extent to which we would like to reengage an actor after she has defected. It allows opportunities for apology and amendment which are more effective than brute punishment. Brute punishment is more easily misunderstood and are coarse-grained in terms of securing suitably amended future behaviour. With the addition of linguistic communication, misunderstandings which might otherwise jeopardise cooperative interactions can be corrected, withstanding signalling errors and ensuring that future interactions involve precisely codified acceptable behaviour. In this way, language might be used to recover one's reputation following a reneged commitment.

[74] The circumstances where this is less likely to hold are where the stakes of the interaction are high and opportunities for future interaction are low. Here, defection pays. Indeed, one might worry about Machiavellian cheaters who only defect when news of their defection will not reach others and so maintain an untarnished reputation. This will be discussed in this next section on deception.

Another kind of fitness advantage derives from the fact that linguistic commitment does not depend on shared activity or social bond creation. Here, rather than the fracturing of a social bond, it is the possible identification of a broken promise that changes *t*+*n* payoffs. At the ultimate level, this is met with potential exclusion. At the proximate level, this is incentivised by fear of ostracism and other such emotions. The fact that shared activity is not needed as a foundation for these commitments to be fulfilled, nor for them to change receiver expectations, means that it is possible to make commitments amongst in-group strangers – agents with whom one has never before interacted and know nothing about.[75] Commitments will alter the agent's fitness consequences as long as there exist future opportunities for beneficial interaction. This will be true of familiar partners as seen in the case of commitment via shared activity. Once we have formed cultural groups, it is likely that there exist future opportunities for interaction simply in virtue of the interconnected nature and practices of the group. Yet, over and above this, reputation sharing *creates* such opportunities through enabling the identification of reliable partners within the group and by opening oneself up to potential praise or criticism for one's actions. In this way, linguistic commitment extends the scope of effective cooperation to new partners.

Linguistic commitment even extends effective cooperation to those about whom we do not have prior information since, once we have language, the sender's future reputation acts as the enforcement mechanism behind their commitment. In other words, all that is required for a linguistic commitment to be credible to an in-group stranger is that one's reputation is on the line. In order for this to be the case, there needs to be potential future interaction with agents who may

[75] Of course, commitment via participation in shared activity also involves strangers at the very first interaction, but trust here is built incrementally and hunting may have been secured by means other than commitment before the rise of redundancy in numbers.

hear of the promisor's behaviours, or else there would be no cost to reneging. So it would need to be the case that an in-group stranger shared the same network of social interaction as the promising agent, even if the two interacting agents know nothing about one another. Initially, this may seem paradoxical. If the agents share a common network, it is likely that they know something about one another. However, this need not be so. Agents can share the same social network yet not have received information about one another before interaction. If this is so, the agent counts as a genuine in-group stranger yet there is a non-negligible chance that this particular instance of the promisor's behaviour would reach her future partners through gossip. This is particularly so when *other* in-group strangers who the agent does not yet know may become potential partners in the future. Of course, the cost of defection will be proportional to the exclusion the reneging promisor faces. Since the chance of news reaching one's future partners is lower in a less-connected social network, the cost of defection is lower, and we would expect that promises are more often broken in this context.

There are two factors which may serve to incentivise promise-keeping despite this. First, promises may be made in a public context. If this is so, agents who renege suffer reputational damage among the people who witness their defection. That is, the agent may be excluded from future interaction with other in-group strangers who observe her behaviour toward a different in-group stranger. Not only this, but those observing in-group strangers may have network ties to the agent's own familiar partners. Though the potential cost of defection for each of these possibilities may be small, together, they might add up to incentivise cooperation. Second, and related, calculating in each instance whether news of an agent's behaviour will reach someone she knows is cognitively demanding. If a Machiavellian strategy is too costly, it might pay to develop the general disposition

of being honest since, on average, it works out in one's favour. Indeed, this explanation might shed light on why, at the proximate level, we are incentivised to follow through on our promises for fear of guilt even when we know no one will hear of our promise-breaking (Vanberg 2008). Selection acts upon our existing emotional repertoire to facilitate cooperation-enhancing behaviours rather than fashioning Machiavellian calculators.

Yet this is not to say that linguistic commitments offer only advantages. There may be disadvantages, too. Implicit commitments can be made unbeknownst to the agent, since all that is required is the change in the receiver's expectations of the sender's cooperation. This would be detrimental where the agent does not act in line with what the receiver has taken her to have signalled (since she will be viewed as a less desirable partner) and beneficial where she does (since she will be viewed as a more desirable partner). There may be reason to believe that agents are not routinely hypocritical so, on balance, our actions will be in accord with our expressed attitudes. However, even if this is so, there does not seem to be any advantage for implicit commitments over explicit ones, given the fact that these commitments can be unwitting. At first pass, it appears agents would always do better by being aware of what commitments they have made. So why do we have unintentional implicit commitments if intentional ones would serve us better? Relatedly, one might ask why we could not just rely on explicit commitments, which are intentional?

The challenge is to explain why both behaviours – implicit and explicit promising – are stable and coexistent. I argue that implicit commitment is stabilised by its secondary evolutionary benefit in reputation sharing. Although the same commitments could be made explicitly, reputation sharing is a beneficial activity which also entails commitments. Reputation sharing is beneficial since it

provides others with information about the state of their social network which is likely to be reciprocated. So implicit commitment is evolutionarily advantageous not only because it involves making a commitment which can advertise our trustworthiness but because it simultaneously shares information about the state of our social network. As a result, unintentional commitments may be a necessary by-product of the benefits of reputation sharing since to keep track of all the ways one has changed other's expectations would involve significant cognitive costs.

However, the two forms of linguistic commitments, explicit and implicit promising, are deployed in different contexts – implicit commitments are made when there are advantages to the sorts of reputation sharing on which it depends and explicit commitments are made when fidelity and clarity concerning an agent's future action is important. This is why, for example, we see explicit commitments more often in cases of joint activity when clear expectations are required to succeed on the task and roles would otherwise be ambiguous. When we use implicit commitments versus explicit commitments will depend on the trade-off between the accuracy of communicating our future course of action and the benefits of sharing and receiving information about the state of our social network. Which type of linguistic commitment is used will also depend on the conversational context. It may be more natural for an agent to make an explicit promise in the case where she does not have a reference action from a third party. Conversely, it will be more natural for her to make an implicit commitment where she is already engaged in a conversation about a third party.

3.7 On deception and exclusion

Shared activity does not suffer dramatically from the problem of cheating since hunter-gatherer societies are intimately connected, compounding the cost of *c.* However, in larger societies, this is not necessarily the case. The cost of exclusion might no longer threaten survival. With this reduced cost of defection, we would expect it to be advantageous to be able to deceptively promise in an environment where promising is widespread. Indeed, since linguistic signals are cheap to fake, one might expect that the signalling system would be undermined by deception. This would be particularly true in the case of interactions with individuals whose reputation is inaccessible or where we might expect that the cost of defection would be low in virtue of fewer future opportunities for interaction. In this section, I argue that, in light of the potential for deception, natural selection has favoured the development of a range of cognitive adaptations which serve to mitigate susceptibility to deception. The result is a population in which there is both cooperation and defection and where the credibility of commitment signals is not undermined.

The conditions for stable cooperation in the face of potential deceptive signalling have been studied from the perspective of game theory. It is known that pre-play signals of cooperation (such as commitment signals) can change the relative size of basins of attraction in Stag Hunt games making cooperation more likely (Skyrms 1996; 2002). For such signalling to remain credible, deceptive signals must be detectable and punished to the extent that they would become disadvantageous from a fitness perspective. So, as with commitment via shared activity, credibility may be ensured by social, if not intrinsic, cost. Note also it is not necessary that the punishment is sufficient to *eliminate* deceptive signalling in order for the signalling system as a whole to remain credible (Skyrms & Barrett 2018). Modelling work suggests that a stable level of deceptive

signalling can co-exist with honest signals in a mixed-strategy equilibrium (Clark & Kimbrough 2017).[76] Indeed, Frank (1988) notes that cooperation based on commitment is likely to result in a population equilibrium containing both cooperators and defectors.

Behnk and colleagues (2018) find that when revelation of dishonest behaviour is guaranteed and is punishable, senders are more likely to be honest. In many circumstances where explicit promising takes place, we would expect that failure to abide by one's promise is directly observable, incentivising honesty. When monitoring levels and detection probability are lower, high levels of honesty can still be maintained as long as punishment is severe and cost-free for the punisher (Behnk et al. 2018). Here, severely sanctioning dishonesty means to reduce the sender's final payoff to an amount that is lower than all other payoffs in the game. So, where $P_R(d)$ is low, honesty can be sustained if c is high and l is low.

It is plausible that social exclusion meets these conditions as a form of punishment, imposing costs on offenders through both negative psychological experiences and reduced access to the benefits of cooperative interaction while involving little to no loss on the part of the punisher (Eisenberger et al. 2003; Williams 2007; Archer & Coyne 2005). This is particularly true in a world with reputation sharing at play, since there will be ample opportunities for future partnering with cooperative individuals. There is also evidence that social exclusion may be chosen as a sanctioning method over and above more costly forms of punishment. While direct punishment strategies such as physical or verbal confrontation are usually effective in amending future behaviour, they are also risky as the punisher may be met with retaliation (Balafoutas et al. 2014;

[76] See Clark and Kimbrough (2017) for an explication of the different punishment regimes.

Nikiforakis 2008). Social exclusion is typically less risky. A study by Molho and colleagues (2020) also found that gossip is often used as a sanctioning method. Since language use was not undermined by deception and the practice of social exchange and promising still persist, cognitive adaptations such as these likely stabilised their effective usage. Indeed, humans often do faithfully carry through their stated commitments (Charness & Dufwenberg 2006; Vanberg 2008; Mischkowski et al. 2018).

Social exclusion is a low-cost form of punishment which directly benefits the agent doing the excluding – the agent is forgoing interaction with an individual who will sucker her. If this is so, there is no difficulty in explaining why one is incentivised to punish. Furthermore, if the dispositions to punish demonstrated in the above empirical evidence serve to keep signals honest through social costs, linguistic commitments will be a reliable means of securing cooperation. One could argue that there are situations in which one has to actively work to exclude a defector from a beneficial outcome, for example, with public goods. In these scenarios, exclusion may be costly to enact. However, as noted in Chapter 2, this kind of costly punishment is not strictly required in order for commitments to operate. Alternatively, if costly punishment is incentivised by means of a conformist tendency, as suggested by Henrich and Boyd (2001), the conformist tendencies which served to stabilise punishment in pre-linguistic societies will become even stronger in the presence of cultural groups with linguistic information-sharing. When the size of the group grows and markers of group membership are required for identification, it is likely that conformism will increase as a means of signalling one's membership.

Linguistic commitments also allow us to commit to behaviours with precise enough detail that consistency (or inconsistency) of our future behaviour with our purported commitment is more easily observed. Whilst defection from cooperation in the case of commitment based on shared activity was only observable in contexts such as a group hunt, defection from cooperation for humans who engage in social gossip is not contingent upon such rare events. It can be observed in a number of small and frequent ways, for example, in not attending the fortnightly ritual dance after criticising Danny for doing so. So, frequent, low-stakes gossip will serve to increase the probability of defection being detected compared to rarer activities.

It is also worth noting that the severity of the threat of deception depends on the relative costs and benefits of honest and deceptive signalling. This in turn depends on the nature of our cooperative enterprises. So deception will not be a threat if $EV(SR1 \ \& \ R2) < EV(SR1 \ \& \ R1)$ since either w is high, z is low, $P_R(d)$ or c is high, and l is low. Indeed, if it is the case that we cooperate in a number of small and frequent ways, for example, in the fair allocation of figs, the benefits of deception on any particular occasion, z, are likely to be low (the agent perhaps gains an extra fig). More frequent interactions also increase the sender's perceived likelihood of defection being detected, that is, $P_R(d)$ is higher. Furthermore, if our interaction is such that we are provided with continual opportunities for low stakes mutually beneficial interaction, the costs of deception, c, are likely to be high, since the agent risks being excluded from the sum of all future resources which are to be divided or mutual gains to be achieved. Social exclusion is a relatively low-cost form of punishment, meaning l is low. Importantly, as we have seen from discussion of the advantages of linguistic commitment, it enables more cooperation. This increases the likelihood that we engage in low stakes interactions. So linguistic commitment *itself alters* the profitability of deceptive

signalling. That is, deception is less profitable when the frequency of low stakes interactions lowers the fitness benefits of defection and raises its cost.

This could lead one to wonder whether reputations are misleading since prudent individuals would not choose to cheat in situations where the payoff, *z,* was low and the perceived risk of detection, $P_R(d)$, was high. Thus, information about an agent's behaviour in these circumstances is not a reliable indicator of their behaviour where the probability of detection is low and the payoff is high. The reason that this ought not to undermine the credibility of reputation-tracking is that cheaters are not perfectly rational. Rationality typically plays only an *indirect* role in motivation, and agents are instead motivated by immediate gains (Frank 1988). The extensive literature on delay discounting supports this.[77] Furthermore, the environment is noisy. It would often be difficult to assess the risk of detection and the potential sanctions imposed. This means, at some point, an agent is likely to give in to the temptation to cheat even when it does not pay to do so. Although the conditional strategy "defect when $P_R(d)$ is low, cooperate otherwise" may fare better in the evolutionary dynamics, it is unlikely that natural selection opted for such a high-rationality route. Indeed, we do not see this evidenced in the behaviour of children. A study by Warneken and Tomasello found that helping behaviour is not contingent upon receiving previous helping or sharing in kind (Warneken & Tomasello 2013) and 2-year-olds help others even if the person they are helping did not solicit help (Warneken 2013). This suggests that we did not evolve to be Machiavellian calculators about cooperation. Indeed, when the frequency of interactions is high and information is readily shared through gossip, adhering to honesty may be the best policy.

[77] For an overview, see de Matta et al. (2012).

It is also important to note that commitments do not operate in isolation – access to another's reputation will bolster the reliability of commitment signals. Indeed, in an environment where people actively monitor social information for its veracity, signalling is risky unless one is committed to following through on one's signalled behaviour since news of one's defection will spread. This means that exclusion is likely not only imposed by the agent to whom one has committed, but others in the social network, compounding the costs of defection and reducing the temptation to deceive. So reputation sharing, as well as various adaptations for the detection and punishment of deceptive signalling, allow for the effective use of linguistic commitment in the face of potential deception. Such adaptations are also aided by the fact that we undertake commitments on a wide range of issues. This provides more opportunities for verifying the reliability of another's commitments, limiting the potential for successful deception.

This is not to say there are no defectors. Indeed, the foregoing implies that the threat of deception is much greater in cooperative interactions which involve higher stakes, and among agents with whom we do not expect to engage in repeated interaction. However, this is not surprising. We are more suspicious of commitments offered by out-group members than we are of those offered by in-group members, and more suspicious of in-group strangers than we are of familiar partners.

Given the mechanisms we have developed to detect and punish deceptive linguistic commitments, I suggest that honest and deceptive commitments exist together in equilibrium. Cheap talk has not undermined the signalling system. Indeed, in congruence with the thesis presented here, Lachmann and colleagues (2001) suggest some signals may be low-cost precisely because they are easier to verify – there is a correlation between ease of verification and cost-free signals. They argue that

the reason costly signals are produced by peacocks for advertising mating quality and cost-free signals are produced by sparrows for advertising fighting quality is because the former's signal reliability is much harder to verify compared to the latter. A peacock's genetic quality is observable only in the distant future and is, even then, stochastic in its effects, while a sparrow's fighting ability is easily verified and deceptive signals are easily punishable. As a result, production costs are not required to ensure reliable signalling (Lachmann et al. 2001). Likewise, I argue that deceptive linguistic information, despite being low-cost to produce, can be kept honest via social, if not intrinsic, costs.

3.8 Expanding our cooperative landscape

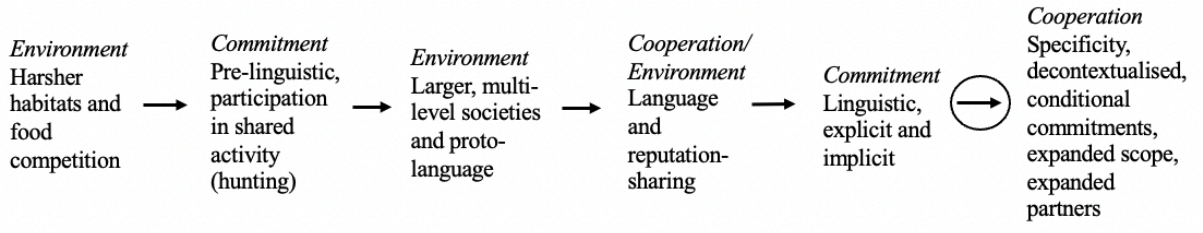

Figure 12: Expanded cooperation as a result of linguistic commitment.

So, I argue, the increased cooperation made possible by commitment via shared activity contributed to the development of larger, multi-level societies and changed the selective environment such that sharing information with others in the group became important. Whether this was for the communication of ecological information in growing group sizes or for the sharing of reputation in response to inability to directly monitor behaviour, such pressures resulted in the evolution of greater communicative capacities and ultimately language. If the sharing of ecological

information in joint tasks such as hunting was the selective driver of the evolution of language, our coevolutionary story is not only about the conditions that needed to be in place for new forms of cooperation to evolve, but it is a *causal* story about what drove that cooperation. In the context of ever increasing and more productive cooperation enabled by information sharing and commitment via shared activity, the ability to make oneself an even more attractive partner through commitment secures further benefits. Thus, the advent of larger groups and language both made possible, and favoured, the evolution of two new forms of commitment, explicit and implicit linguistic promising.

Crucially, the evolution of linguistic commitment permitted us to expand cooperation further than was possible with commitment based on shared activity. In particular, linguistic commitment applies to more contexts, those captured by Prisoner's Dilemma and Trust games. It allows us to extend cooperative interaction to in-group strangers with whom we might not have opportunities for shared activity, but with whom we might accrue fitness benefits from repeated interaction. It enables more precise communication, resulting in more grounded expectations of behaviour. It also permits conditional promises and promises whose consequences are specified. If an agent offers tangible collateral, her partner may be more likely to believe her promise, making an action which was not otherwise subject to a commitment to become so. Furthermore, via linguistic commitment, we expand the range of matters on which we can undertake commitments since language is flexible enough to matter for cooperation in new contexts and can be used to communicate commitments concerning issues which are spatially and temporally remote from the current environment. This also provides more opportunities for assessment of a partner's reliability

since it provides more scenarios in which they can demonstrate themselves as trustworthy or untrustworthy, thereby aiding in correlated interaction among cooperators.

Thus, we have a coevolutionary link between commitment and cooperation – the cooperation enabled by earlier small-scale collaboration and ecological pressures which resulted in the formation of larger groups and the development of language within such groups led to two new forms of commitment – explicit and implicit linguistic commitment. These new forms of commitment made possible new forms of cooperation through the benefits of expanding our range of commitment and the scope of our potential partners. The evolution of commitment therefore had a profound impact on the evolution of human cooperation. To elucidate the effect of linguistic commitment, it is useful to note that some cooperative activities could not have been achieved without the ability to communicate about specific divisions of labour, expectations and abstract events. Consider, for example, collective action problems such as building projects. Here, the ability to abstractly refer and precision about the division of labour in financial and physical investment would greatly aid in achieving mutual benefit. However, communication alone is insufficient for agents to trust one another when there is an opportunity to free-ride. What is required is the ability to make a credible commitment to investment – that is, the ability to linguistically promise.

It is also important to note that, as our cooperative possibilities extend, we might require cooperation to be temporally extended. That is, cooperation may involve future actions rather than, for example, presently hunting a stag. Such scenarios might include acquiring resources for a future building project; bringing a particular dish to a dinner party; collecting a child from daycare.

If this is so, we require the ability to commit concerning future events and language enables us to do this by placing our reputation on the line. Sterelny (2012) notes the very different trust requirements of cooperation based on immediate return mutualism and cooperation based on reciprocal exchange. In cases of reciprocation, a partner may repay cooperation at a later date and in a different form. If this is the case, it is more difficult to ensure successful cooperation as cooperation requires trust in the other's reciprocation. I argue that this trust would be advanced by the connection between commitment and reputation with the advent of linguistic commitment. If an agent is able to credibly commit and follow through on her commitment, she will be more trustworthy in cooperative interactions that rely on later reciprocation. News of her character will also spread to others in the network. So, as the nature of our cooperative enterprises change from being those which involve immediate returns to those which involve delayed returns, linguistic commitment serves to secure greater correlated interaction. The advent of linguistic commitment thus smooths this transition to a riskier form of cooperation. In the next chapter, we will see how norm externalisation allows for even more powerful commitments.

Chapter 4: Moralised commitment

The advent of norm externalisation furthers the scope and power of linguistic commitment. To say that a norm is externalised is to say that it is experienced as imposed on us from the outside and exacting a demand on all, regardless of their group (Stanford 2018a). For ease in this chapter, I will call commitments which refer to externalised norms "moralised commitments", though, as will be shown shortly, the category of the externalised does not perfectly correspond to all and only moral norms. While placing the emergence of externalised norms in time is near impossible, I will argue that it has deep ties to our previous practices of commitment and cooperation and that it stands in a coevolutionary relationship to what came before it. Our previous practices of commitment and cooperation provided the cognitive precursors to externalised norms in the development of our perspective-taking capacities and the affective mechanisms that motivate helping behaviour. Not only this, but hunting and linguistic commitment contributed to growth in the size and complexity of the group, which itself creates selection pressure for the emergence of externalised norms as a means of correlating interaction. We will see that commitments which refer to externalised norms – moralised commitments – can come in different forms and that they offer a number of fitness advantages over non-moralised commitments.

Note that it will not matter to my thesis whether moralised commitment succeeded or preceded linguistic commitment which was not moralised. My thesis is that new forms of commitment coevolved with new forms of cooperation, and this will be true whether some of these coevolutionary processes were happening simultaneously. What matters is that our previous forms of commitment and cooperation (hunting and the development of larger groups with reputation sharing practices), led to the evolution of more effective commitments to facilitate correlated

interaction in increasingly interconnected worlds. In this chapter, we add the following to our coevolutionary story: commitment via shared activity set the stage for the development of externalised norms which, along with language and reputation sharing, provided the resources to extend the scope and power of our commitments (Figure 13).

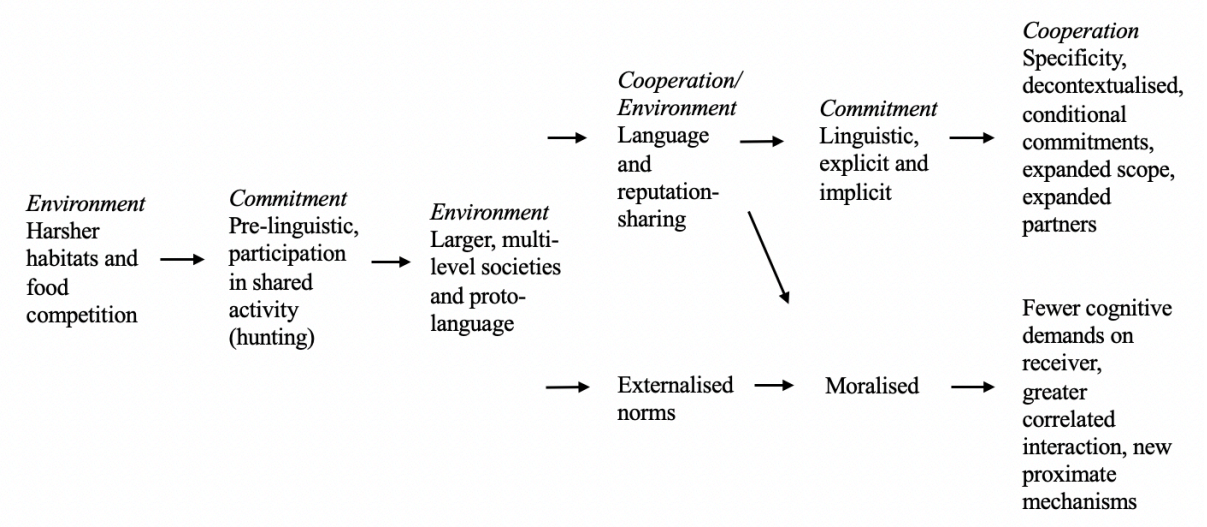

Figure 13: The coevolution of commitment and cooperation.

4.1 Externalisation and moral norms

Stanford (2018a) introduced the concept of externalisation in discussion of what he took to be the distinctive phenomenology of moral norms. Cross-cultural research has suggested that we distinguish moral from conventional violations at a young age (Turiel & Nucci 1978; Turiel 1979; Nucci et al. 1983; Smetana & Braeges 1990; Nucci & Turiel 1993; Nucci 2001; Huebner et al. 2010).[78] Not only this, but these violations influence human behaviour differently. Skitka and colleagues (2005) find that while there is not a clear systematic relationship between non-moral

[78] Where "conventions are part of constitutive systems and are shared behaviors (uniformities, rules) whose meanings are defined by the constituted system in which they are embedded" (Turiel et al. 1987: 169).

disagreements and desirability for social interaction when we control for attitude strength, there *is* a systematic correlation between disagreement on *moral* issues and preferred social and physical distance.[79] Attitudes treated as rooted in moral convictions influenced interpersonal measures in ways beyond non-moral attitudes with analogous "attitudinal strength". These interpersonal measures included preference for social or physical distance; intolerance for attitudinal difference in both intimate and distant relationships; lower reported good will and lower cooperativeness in solving group tasks. It was found that attitude strength alone was generally unassociated with preference for social distance from dissimilar others (correlations were found in 3 out of 24 cases tested), and where this did occur, the preference did not generalise across multiple issue domains. Yet disagreement on moral issues was a good predictor of preferred social and physical distance.

Furthermore, preschool children demonstrate higher levels of moral objectivism than 9-year-olds and adults, suggesting the externalised nature of moral norms is a pervasive feature of human experience, perhaps preceding explicit teaching (Schmidt et al. 2017). This is in contrast to other judgments. A study conducted by Nichols and Folds-Bennett (2003) showed that children from 4- to 6-years-old would often concede, when faced with disagreement, that grapes were yummy only "for some people" but would not exhibit such concessions in the case of morally good or bad actions, such as monkeys helping one another. This study shows that taste judgments are not typically externalised in the same way as moral judgement.

[79] The measure of social distance used is an adaptation of the measures developed by Byrnes and Kiger (1988) and Crandall (1991). Participants were asked the indicate the extent to which they agreed or disagreed with different completions of the sentence: "I would be happy to have someone who did not share my views on (their identified most important issue)…". Completions were "as President of the U.S.," "as Governor of my state," "as a neighbor," "to come and work at the same place I do," "as a room mate," "to marry into my family," "as someone I would personally date," "as my personal physician," "as a close personal friend," "as the owner of a store or restaurant I frequent," "as the teacher of my children," and "as my spiritual advisor." Physical distance was measured by placement of chair from the dissimilar other.

Of course, there may be differences in the particular behaviours which are viewed as moral or matters of societal convention in different cultures. Moral norms are sometimes thought to be those norms which concern issues of welfare, justice or rights (Nichols 2004). However, some actions which do not fall under these categories (such as cleaning one's toilet with the national flag) can also be moralised (Haidt et al. 1993; Nichols 2004). Those societies that are heavily influenced by religions may include concepts such as sexual purity under the purview of morality (Vasquez et al. 2001). Indeed, evidence suggests that Indian and Muslim societies do not draw any distinction between moral and non-moral norms (Machery 2012). It has also been observed that etiquette transgressions trigger responses typical of moral transgression in American children, for example, strong feelings of disgust (Nichols 2004). As such, one may question whether there is truly a moral-conventional divide, and thus whether such externalisation is in fact uniquely a feature of moral norms.

However, whether or not we identify externalisation with only moral norms is not crucial to my account. The important part of this account is that we externalise some norms and that we signal attitudes about these norms which can be taken as commitments, whether the boundary of what counts as externalised is clear or whether it exists on a continuum. Experiments have shown that there exist judgments whereupon disagreement between parties is taken to be an indication that one party is *mistaken*, supporting the idea that at least some judgments are externalised (Goodwin & Darley 2008). That one party is thought to be mistaken reveals that these judgments are treated as an objective disagreement rather than a subjective disagreement since, in the latter, two people could differ in opinion but both be right. It is these judgments which are the focus of my inquiry

and provide the basis for an extension of linguistic commitment. Nevertheless, while we are concerned with *externalised* norms rather than moral norms, the latter will act as a good proxy for the former, since many typical moral norms are "unconditionally obligatory, generalizable, and impersonal" – that is, externalised (Turiel et al. 1987: 169–70). As there is ample psychological evidence concerning moral attitudes and behaviour but little conducted solely on externalised norms, we will use this empirical data as a proxy for understanding attitudes and behaviour concerning externalised norms.

4.2 The emergence of externalised norms

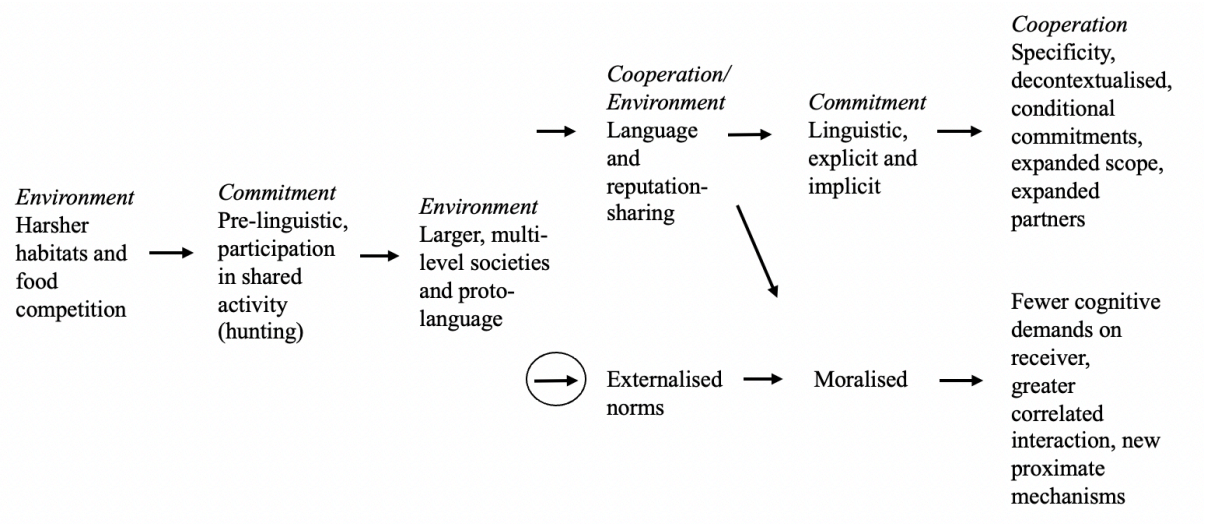

Figure 14: The emergence of externalised norms.

How and why did norm externalisation emerge? I argue that externalised norms coevolved with previous practices of cooperation and commitment in two ways. First, I will argue that moral cognition, of which externalisation is typically a part, has its foundations in the prosocial emotions and perspective-taking capacities which coevolved with previous forms of commitment. Second, I argue that externalisation is selected for as a mechanism of partner choice in larger and more

complex societies. Harking back to a distinction made in Chapter 3, I will make two kinds of claims: one concerns the cognitive and emotional *precursors* to the emergence of externalised norms and the other concerns the *selection pressure*. The first of these coevolutionary claims is not causal, but rather suggests that commitment had a role in providing the precursors that needed to be in place for a new form of cooperation to evolve (correlated interaction based on externalised norms). That is, commitment via shared activity did not select for externalisation but it created the conditions that made externalisation possible – the cognitive foundations of moral thought. The second of these coevolutionary claims is causal. I suggest that the advent of larger societies with more complex cooperative interactions resulting from hunting and language selects for externalisation as a means to facilitate correlated interaction.

4.2.1 The precursors to moral cognition

First, I will establish the claim that our previous forms of commitment coevolve with the cognitive and emotional precursors to externalisation. Sterelny boldly writes, "prosocial and commitment emotions evolved before moral cognition; they made possible the cooperation and cultural learning that prepare the evolution of explicit normative thought" (Sterelny 2012: 166). According to Sterelny, moral norms grow from our dispositions to respond positively to cooperative interactions and the emotions involved in such interactions make these events salient to us. He argues we not only respond positively to kindness, but we are *aware* of that positive response and *convert* this into a normative judgment.[80] Sterelny (2012) does not detail *how* exactly these prosocial emotions come to be "converted" into normative judgments. I hold we can bridge this gap by appeal to the

[80] Of course, moral thought can, in turn, come to shape our emotional responses.

work of Goldman (1993), Gordon (1995) and Nichols (2004) on the role of perspective-taking, empathy and affective mechanisms in moral cognition.

The Piagetian tradition sees moral understanding as fundamentally involving the capacity for perspective-taking (Piaget 1932; Kohlberg 1984; Selman 1980; Damon 1977). Indeed, even philosophers like Rousseau see empathy as the source of moral principles. More recent work in the field of moral psychology ties empathy to perspective-taking. Goldman (1993: 351) writes, "paradigm cases of empathy… consist first of taking the perspective of another person, that is, imaginatively assuming one or more of the other person's mental states." Deigh (1995: 758) argues that grasping right and wrong depends on taking another's perspective and imagining feelings of frustration or anger.[81] As we have discussed in Chapter 3, a key component in the development of such perspectival representations were the requirements of effective action in shared activity, such as hunting, foraging or alloparenting. Gordon (1995) codifies how this turns into a moral judgment, which is agent-neutral in the way we expect of externalised judgments, by suggesting: "Imagine being in *X*'s situation, once with the further adjustments required to imagine being *X* in *X*'s situation and once without these adjustments. If your response is the same in each case, approve *X*'s conduct; if not, disapprove" (Gordon 1995: 741). We thus see how the perspective-taking capacities developed in situations of joint activity are an important step in coming to make an agent-neutral judgment and, if something like this Piagetian account is correct, the development of a sense of moral right and wrong.

[81] Nichols (2004) believes the mind-reading capabilities required for moral thought are more basic than in the Piagetian tradition, and only involve representation of distress.

Even if we were not to accept Gordon's account, there are many others who believe the roots of moral feelings lie in perspective-taking. Consider, for example, Tomasello's (2016) views on the origins of human morality or Darwall's (2006) second-personal view of morality. Tomasello believes that the origins of morality lie in the recognition of the interdependence we have with others. He writes that this recognition factors into our decision making so: "(1) that helping partners and compatriots whenever possible is the right thing to do, (2) that others are equally as real and deserving as themselves (and this same recognition may be expected in return), and (3) that a "we" created by a social commitment makes legitimate decisions for the self and valued others, which creates legitimate obligations among persons with moral identities in moral communities" (Tomasello 2016: 160). It is again important to his account that this derives from "shared intentional activities" (reminiscent of the shared activities of Chapter 2) where the concept of "we" is in control of the "you" and the "me". Participation in such shared activities involve the participants taking on another's perspective and trying to manipulate their perspective using cooperative communication. Later on in Tomasello's account, morality takes on an agent-neutral perspective. Darwall's (2006) position also crucially involves taking second- and third-personal perspectives in relation to the claims we make on one another's conduct and will. So I am not tied to any one account of the origin of morality – only the less controversial claim that perspective-taking is an important part of moral cognition. Importantly for my purposes, it is in the context of joint activity that we see such capacities develop.

Indeed, the hypothesis that perspective-taking is important to the evolution of moral feelings would also explain why great apes do not evolve moral norms, even though they have prosocial

motivations (Tomasello 2016; Warneken & Tomasello 2006).[82] Children, but not apes, exhibit the capacity for "joint intentionality", in particular, the kinds of socially recursive inferences, social self-monitoring and recognition of self-other equivalence that precedes normative thought. Children as young as 14 months will proactively help unfamiliar adults with fetching objects and, at 18 months, they will proactively help to open doors and stack objects when they can see that another cannot complete the task (Warneken & Tomasello 2007; Warneken & Tomasello 2006).

In contrast, it has been found that while chimps demonstrate the capacity to take on another's perspective in competitive tasks, they do not in cooperative tasks. A study by Hare and colleagues (2000) showed that conspecific subordinates would choose pieces of food hidden from the view of a dominant chimpanzee instead of food visible to them. Of all the food obtained, 83 per cent came from the subordinate's own cage, visible only to them. Furthermore, on seven occasions, the subordinate waited until the dominant had moved away from the doorway connecting their cages in order to obtain this food. This experimental evidence suggests that chimpanzees understand what is visible to others, operating with at least some theory of mind.[83] However, Hare and Tomasello (2004) find chimpanzees are unable to understand when humans attempt to *altruistically* guide their attention toward food contained in a box. None of the six chimpanzees checked the box for food, where three out of six of them reached for the box when the human had reached for it more aggressively, as if in competition. In light of this evidence, it is hypothesised that chimpanzees lack mind-reading abilities in situations unlike those in which primate cognition

[82] Or they possess what one might call a morality of "sympathy" rather than a morality of "fairness". See Tomasello (2016). Note that some argue apes do engage in normative thought. See, for example, de Waal (1996).
[83] A follow up study found subordinates adjusted their behaviour when the dominant individual who witnessed the hiding of food was replaced with another dominant individual who had not witnessed it, demonstrating the ability for chimpanzees to keep track of who has seen what (Hare et al. 2001). See also Seyfarth and Cheney (2012) for evidence of a theory of mind in chimpanzee alarm calls.

evolved (Hare & Tomasello 2004).[84] Indeed, it is likely that commitment via shared activity characteristic of the hominin lineage plays a large role in the development of such capabilities in a cooperative context.

Nichols (2004) further argues that moral motivations depend on a particular *affective mechanism* that is activated by suffering in others. I argue that this affective mechanism is strengthened by social bonding in shared activities. Why do we need an affective response alongside mind-reading abilities, and what is the character of this affective response? Nichols (2004) appeals to evidence about psychopaths to show that perspective-taking is not enough to explain moral psychology and motivation. These are individuals who have the capacity for mind-reading but are not motivated to respond to the distress cues of others. On the other side of the coin, some autistic children who have difficulty with perspective taking are responsive to distress in others (Blair 1999). So what is needed to explain helping behaviour is either, according to Nichols (2004), a distinctive emotion of sympathy or of empathy. Sympathy does not require that we feel the same emotional distress as the affected agent, but is posited by some as a distinctive, and motivationally powerful, emotion associated with its own physiological characteristics (Eisenberg & Fabes 1990). Alternatively, it is possible that emotional contagion results in *empathy*. Emotional contagion is demonstrated even in newborn infants who are more disposed to cry when hearing the sound of another newborn cry (Nichols 2004). Feeling similar feelings of distress to the other agent will likely motivate moral helping behaviour. The idea is that if distress is mentally represented in the mind of the witnesser, agents will *help* the other in need rather than escape the situation.

[84] One might, of course, question the validity of this conclusion considering the unnatural set-up.

It is found that such distress representation and helping behaviour exists early in ontogeny. Attribution of pain to others is seen in studies with children just 2-years-old (Wellman et al. 1995). Not only this, but it is sometimes accompanied by comforting (Nichols et al. 2009). I hold that commitment via shared activity creates selective pressure for experiencing the affective states of others. This is because these forms of cooperation made us further perceptually tuned to our emotions and the emotional responses of others. As was argued in Chapter 2, hunting creates social bonds more powerful than bonding established in calmer waters, an effect termed "arousal", evidenced by studies in war. Other forms of shared activity were also important. Alloparenting engages our dopamine and oxytocin-related reward systems in a similar manner to parental care. Indeed, our brains our wired to register signals of infants' needs, our endocrine systems induce a rapid response, and our reward systems reinforce these behaviours (Hrdy 2009). As such, these shared activities nurture affective mechanisms that respond to the distress cues of other conspecifics. If this account of moral motivation is correct, shared activity played a role in the development of moral cognition, not only through its importance for perspective-taking, but also in its strengthening of affective responses which motivate helping behaviour.

Of course, prosocial emotions like empathy evolved before hunting and alloparenting – their earliest forms likely in kinship relations. I am not suggesting moral cognition *could not have* evolved without these shared activities, rather, that these cooperative activities further *built upon* our tendencies to behave prosocially and so favoured the development of moral thought. Now that we see the connection between previous commitment activities and the preconditions for the emergence of agent-neutral moral cognition, we must ask *why* moral norms developed and, in

particular, why they frequently take on an externalised character. To understand the psychological phenomenon of externalisation, we must look at the role of partner choice.

4.2.2 The selective environment for externalisation

In this section, I argue that externalisation was selected for in the larger and multi-level societies which were enabled by commitment via shared activity. Recall that hunting, among other factors, permitted growth of the multi-level society due to increased means of subsistence, a division of labour and specialisation which allowed greater reproduction and greater group security, among other factors. Language also allowed us to engage in larger-scale collective action problems with the ability to communicate about specific divisions of labour. It was in this context of these larger and more complex cooperative societies with new interactive partners that we faced greater pressure for effective partner choice mechanisms. In larger and more complex societies, opportunities to stabilise cooperation through partner control are not as effective, since one may not interact with the same agent twice (and therefore benefit from their changed disposition). This makes it even more important to choose wisely. In Chapter 2 we saw that commitment via shared activity is a means of ensuring wise partner selection. It allows receivers of commitment signals to identify other cooperators and allows senders to advertise their trustworthiness. However, shared activities might not arise for all members. In light of this, we would do best to make use of an additional partner selection mechanism.

Stanford (2018b) argues that experiencing some norms as externalised will aid correlated interaction since it both motivates an agent to conform to the norm as well as to use the conformity

of others as a means of choosing partners. The agent sees the norm as both applicable to herself and to others and prefers interacting with those agents who share her views. Given its dual functions of advertising one's trustworthiness and of protecting oneself from exploitation, externalisation offers a solution to the problem of partner choice in disconnected groups and it does so in a more robust way that reputation sharing. This is because, though reputation is a means by which third party information can be used to discriminate amongst partners, it does not allow us to directly *advertise* our trustworthiness to others since it does not (by itself and without commitment) change our motivations to behave in particular ways. With reputation sharing, our attractiveness is only a feature of what others have said about us or what is observable from past plays.

Indeed, Sperber and Baumard (2012) argue that, given the wide variety of cooperative scenarios we face and the variety of human motivations, cooperativeness could not be effectively assessed without making inferences about another's dispositions and mental states. They argue that our intuitive psychology in partner choice depends on two components, "a first component that infers beliefs from desires and actions, infers desires from beliefs and actions, and predicts actions given beliefs and desires; and a second component that, given instances of beliefs-desires-actions patterns, infers psychological dispositions or 'character'" (Sperber & Baumard 2012: 507). This inferred disposition determines how likely we believe it to be that an agent will behave in similar manner in the future. What is important to note is that, if this account of our partner assessment is correct, we face selection pressure to *evidence* our disposition.

Yet to appear moral and cooperate only when one's behaviour is observed, and to otherwise be immoral and defect, demands cognitive effort and is a risky strategy. Indeed, many studies find that complete control over one's image is difficult as predicted behaviour is often drawn from indirect cues (Ambady & Rosenthal 1992; Brown 2003). If this is the case, it may be more cost-effective to *be* moral rather than to merely *appear* moral (Frank 1988; Sperber & Baumard 2012; Handfield et al. 2018). One way to achieve this is to take moral norms concerning cooperation to be external demands, which can then be advertised to others. Thus, externalisation serves to make an agent more trustworthy than her Machiavellian counterpart, who defects when it pays to do so.

We cannot stop here, though. If all that was needed was reliable dispositions for trustworthiness, why would natural selection not have opted for simply instilling in us a *strong desire* to act morally, rather than developing the complex phenomenology of externalisation? Stanford (2018a) provides a response. He argues that the reason we experience some norms as externalised is that this establishes a *connection* between an agent's own motivation to adhere to a certain norm and her means of choosing reliable partners – those who share her views.[85] Externalisation is an efficient way to ensure that these otherwise distinct motivations are systematically tied to one another. Though a sufficiently strong desire would suffice for either one of these motivations, to systematically connect the motivations of an agent and her assessments of her partners, even as the *content* of the norms change, significantly improves correlated interaction. If an agent's motivation to engage in moral behaviours were experienced as only a strong desire, this does not necessarily entail that she demands conformity to this behaviour from others, and by not doing so,

[85] Note, whether not externalised norms are *necessary* for cooperation does not matter – what matters is that it is a mechanism by which we *actually do* correlate interaction, regardless of whether other means suffice.

she would leave herself open to exploitation by those who do not possess the same desire (Stanford 2018a).

We have, then, another coevolutionary step in our account of commitment and cooperation. The perspective-taking capacities that were developed, and affective mechanisms that were strengthened, as a result of shared activities laid the cognitive and emotional preconditions for the emergence of moral norms. Commitment activities also led to larger, multi-level groups, increasing the importance of partner choice for sustaining cooperation. However, in this environment, previous commitment mechanisms become less effective, as opportunities for shared activity would not have arisen among all members. Externalisation of one's moral norms might thus have served as a means of advertising one's trustworthiness in this newly expanded world. Not only this, but externalisation simultaneously protects an agent from exploitation by imposing conformity to one's moral views on others. It is therefore a more effective correlation device than reputation sharing.

This is not to suggest that there were no other important features in the emergence of externalisation, there were likely many. My concern, however, is with how previous forms of commitment and cooperation contributed to the selective environment for new forms of cooperation and commitment to evolve. The first of the coevolutionary claims presented was about the cognitive and emotional preconditions that needed to be in place for commitment to result in a new form of correlated interaction, and concerns *how* externalised norms came to be. The second of these coevolutionary claims is a causal claim about *why* externalised norms came to be – due to the benefits of wise partner choice. Of course, I also do not mean to suggest that externalisation

only has positive consequences for cooperation. Externalisation of norms is also a source of conflict. Communities are normatively diverse, and externalising norms reduces mutual tolerance, increasing conflict costs. The claim is that the benefit of correlated interaction may outweigh the potential for fitness-reducing conflict. In the next section, we will see how norm-externalisation aids in the making of commitments.

4.3 Moralised commitment

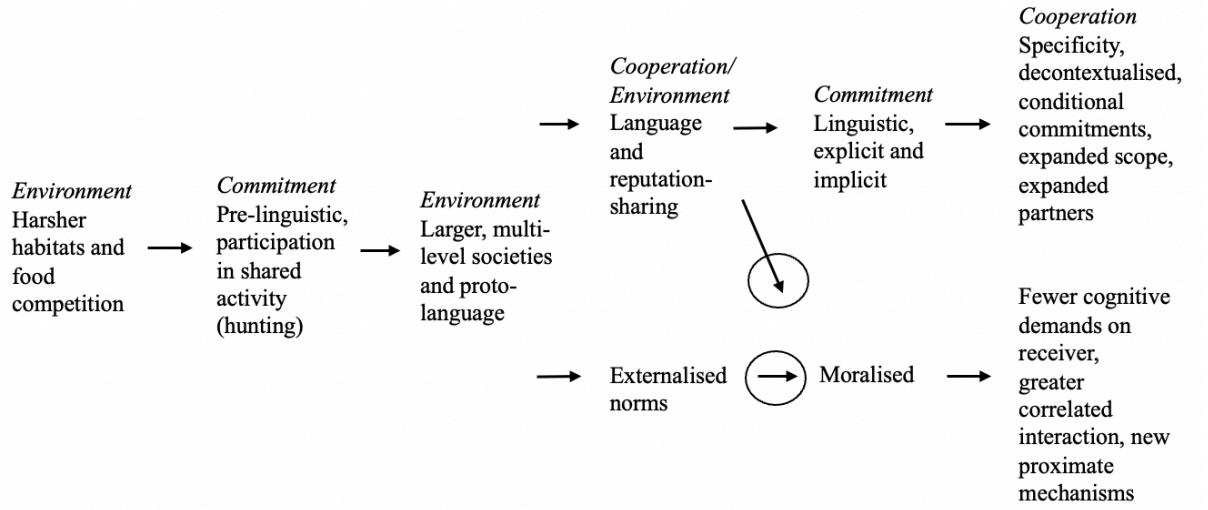

Figure 15: The emergence of moralised commitment.

Moralised commitments are those which refer to externalised norms. These can come in many forms. One can directly assert "doing *X* is wrong", thereby implicitly committing to not do *X.* One can tell a story or fable which expresses a moral attitude, thereby implicitly committing to acting in line with the moral of the story. Finally, one can express attitudes concerning the behaviours of third parties toward externalised norms, by either praising or criticising them. In all cases, the feature that makes these utterances a commitment is that externalised norms are taken to apply to

the utterer of the norm themselves since they are considered universally applicable. To elucidate, if Sandy tells Betty "stealing is wrong" and subsequently chooses to steal, Betty believes her to be a hypocrite and therefore a less desirable interactive partner. Since this utterance changes Sandy's relative payoffs for stealing as it risks potential exclusion, and changes Betty's expectations of Sandy's future action, it constitutes a commitment. It is important to note that on my account, analogously to the case from the previous chapter, the sender of the commitment signal need not take the norm to be externalised in order to fare better if she follows through on her commitment. It is sufficient that the receiver believe that the sender takes the norm to be externalised and is prepared to exclude her on the basis of hypocrisy. This is because what changes the sender's fitness consequences on my account of commitment thus far is potential exclusion from future interaction, and an agent risks exclusion insofar as her *partner* believes her to be a hypocrite.

In terms of our definition from Chapter 1, *a commitment is a pre-play signal in a strategic interaction taken at time t, that increases the sender's relative payoff for carrying through option X at time t+n, and increases the receiver's probability of the sender carrying through option X.* In the case of moralised commitment, the move that the agent takes at time *t* is her expression of an externalised norm, either directly, or indirectly through gossip or storytelling. Option *X* – the option that the agent ensures she carries through with greater probability – is the behaviour which adheres to the externalised norm. Time *t+n* indicates when the interaction concerning the norm takes place. The agent would need to be in a situation in which the externalised norm she has invoked is applicable to her current action. If Sandy criticises Danny for neglecting his children but she does not yet have any children of her own, time *t+n* is in the distant future or may never arise. As with the implicit commitments of Chapter 3, these commitments would need to be

sensitive to the memory of the agents and would be most effective when they are uttered in the context of a relevant interaction or when the time between *t* and *t+n* is short.

As with our previous forms of commitment, the pre-play signal alters the payoffs of downstream choices, incentivising acting in line with one's signal. The incentive to cooperate depends on an increased cost of defection as a result of the pre-play signal. This cost of defection takes the form of either potential exclusion by the partner one has defected on in repeated interaction or exclusion from others as a consequence of a tarnished reputation. As in the previous chapter, the one-shot extensive form game representation in Chapter 1 captures the change in subjective payoffs as a result of commitment signalling while reduced payoffs in the evolutionary game are captured in the agent's reduced opportunities for future interaction with others.

In order to show that moralised assertions are commitments, we must show that there are future opportunities for beneficial interaction and that reneging on one's commitment is punished. As with linguistic commitment, opportunities for future beneficial interaction exist in virtue of the interconnected practices of the group and in virtue of knowledge of potential partners' trustworthiness via reputation sharing. Regarding punishment, it is well-documented that social exclusion is an effectively employed means of punishment for norm violation (Ditrich & Sassenberg 2016; Kerr et al. 2009; Rudert et al. 2020; Molho et al. 2020). Rudert and colleagues (2020) find evidence that ostracism of a target is viewed as legitimate when the target's behaviour violates social norms but many of the examples of norms they cite are typically *moral* (thus, likely *externalised*). These include lying, cheating and lack of contribution to public goods. The proximate mechanisms by which norm-violation is punished include powerful moral emotions.

Indeed, Hopfensitz and Reuben (2009) find that not only does anger play a role in the facilitation of punishment and resultant cooperation, but also guilt, which mediates the desire for retaliation in response to such punishment. In addition to this, we want to know that norm-violation is punished *more severely* when preceded by moral assertions than when it is not.

The adaptations we have developed for detection and punishment of hypocrisy are evidenced in the widespread feelings of *schadenfreude* (pleasure derived from another's misfortune) at the exposure of someone's immoral behaviour, amplified when the person is found to be doing the same action as they were criticising others for (Powell & Smith 2013). Jordan and colleagues (2017) find compelling evidence that hypocrites are perceived more negatively than individuals who engage in the same transgressive behaviour without engaging in prior gossip; more negatively than direct liars who asserted that they did not engage in such behaviour but yet did so; and more negatively than "honest-hypocrites" who admitted their transgressions after engaging in moral gossip. Furthermore, that honest-hypocrites were not judged so severely suggests that what explains our negative attitude toward hypocrites is not their unpredictability or weak-willed nature, but that they have deceptively signalled their attitude. These adaptations for feelings of *schadenfreude* or outrage at moral hypocrisy show that defection preceded by false signalling is more severely punished than that which is not, meaning commitments increase the stakes of defection.

In much of the forthcoming discussion, I will focus on moralised commitment in the form of gossip rather than direct assertion of the form "doing *X* is wrong" since Jordan and colleagues (2017) find that gossip functions as a stronger signal of one's own moral goodness than direct assertions of

one's moral values.[86] That is, it more robustly changes receiver expectations of the sender's future actions. Such signals will take the form, "it's wrong/bad that Danny did *X*". Indeed, there is evidence that we frequently engage in such moral gossip (Dunbar 2004). Bergmann notes that the subject matter of gossip commonly concerns "character flaws, discrepancies between actual behaviour and moral claims… socially unaccepted modes of behaviour" (Bergmann 1993: 15). Fernandes and colleagues (2017) find that the motivation to gossip is stronger when the behaviour of the target was perceived to be in violation of the gossiper's own moral foundations. Insofar as moral norms are generally considered to be externalised, this supports the thesis that such gossip statements function as signals of commitment to future behaviours. Nonetheless, the aforementioned diversity of ways of making moralised commitments is advantageous since we will naturally make commitments in the course of passing on information that is useful in other respects, for example, in moral instruction of the form "doing *X* is wrong" or fables which seek to demonstrate this.

4.4 The advantages of moralised commitment

Moralised commitments offer a number of fitness advantages compared to non-moralised commitments which help to explain their emergence as a means of securing correlated interaction. First, note that the implicit commitments of the previous chapter only applied where the receiver has reason to believe that the sender is part of the same cultural group and occupies the same social role as the person about whom they are gossiping. If Sandy says to Betty, "it is terrible that Danny did not attend the fortnightly ritual dance", this would not count as a commitment if Betty did not

[86] The authors suggest this may be because people monitor social information for its veracity and overt self-promotion may therefore backfire.

believe that Sandy was part of the same tribe and therefore would be a hypocrite if she did not attend the fortnightly ritual dance, nor would it count as a commitment if only men needed to attend the fortnightly ritual dance. This is because it is only in virtue of changed receiver expectations of the sender's actions that the sender's payoffs change – she risks exclusion on the basis of hypocrisy if the receiver takes her utterance to imply something about her future action. Here, the receiver would need to know something about the agent to make such a judgment. In particular, that Sandy is of the same tribe and of the appropriate social role to be obligated to attend the dance.

In contrast, with moralised commitments such as the utterance "it is wrong that Danny stole all the figs", Betty needs less information about Sandy to hold her to be making a commitment. If Sandy and Danny do not belong to the same cultural subgroup (foraging party) and occupy the same social roles (distributor of fruit), what about Sandy's criticism of Danny signals something meaningful about her attitudes? Sandy's declarations of her attitudes are still informative for Betty due to externalisation. If Betty believes that Sandy takes the norm she is espousing to apply equally to all, including herself, Betty may be prepared to exclude Sandy on the basis of hypocrisy if she acts in contradiction to her signal. The fact that this invokes an externalised norm screens off all other features of the agent as irrelevant to understanding them as making a commitment. The commitment is in a sense, more automatic, which is useful in larger networks where agents might interact with someone they do not previously know anything about. So commitments become easier to advertise for the sender and easier to discern for the receiver, aiding wiser partner choice. Of course, this depends on a readiness to attribute moral judgment to the sender (over attributing

cultural background and social role), but this is plausible given how seriously we take moralised language.

This is not to say that non-moralised implicit commitments do not also provide information about the sender's future course of action. As we have seen in Chapter 3, praise or criticism concerning religious or cultural norms will also carry such information and change the receiver's expectations of the sender's future action. The point is rather that externalised norms provide a particularly good way of doing this since they abstract away from many of the details of the signaller and therefore change receiver expectations of sender cooperation more reliably.

However, this again raises the question of why an agent would not simply commit explicitly, rather than via gossip about another, for example, by saying "I will carry that log with you". This would equally screen off the relevance of beliefs about the sender's cultural background and social role. The answer is that agents could just as well explicitly promise, but moralised commitments provide a second advantage. There is an important difference between saying "I will carry that log with you" and "not carrying that log with you would be *wrong*" or "it is *wrong* that Danny did not carry that log with you". In the former case, the agent advertises her trustworthiness as a potential partner. In the latter two cases, the agent not only advertises her trustworthiness but, by invoking moral language usually applicable to externalised norms, simultaneously reveals to the receiver that she is willing to exclude others for not contributing to carrying logs as well, since this norm applies not only to herself but the person about whom she is gossiping *and* the receiver of the commitment signal. In this way, she advertises her expectations of similar behaviour on the part of her interlocutor. This is beneficial for the sender insofar as it protects her from interaction with

those who are likely to behave differently to her since she has revealed that she will not interact with them in the future if they do so. It is easy to see how this would be useful for correlating interaction when interacting with outgroup members where we might not share the same expectations about others' behaviours. However, in the ancestral environment, we would expect there to be more intragroup interaction than intergroup interaction.

Yet moralised commitments also serve to advertise the sender's expectations of others more clearly *within* one's group than non-moralised commitment. We have seen that moralised commitment, compared to implicit commitment on the basis of gossip, lessens the demands on a receiver of making inferences about the sender. Earlier, we established this was the case for making inferences about the sender's future actions, but it is also true of making inferences about the sender's expectations of others. Why do expressions concerning cultural norms not carry the same information about the sender's expectations of others? They carry only partial information about this – the features that make it such that the agent would exclude the third party about whom they are gossiping might not apply to their current interlocuter, since the third party might not share the same cultural background and social role as the receiver of the commitment signal. That is, when we express attitudes about cultural norms, it is unclear whether the endorsement or rejection is contingent on particularities of the person, their role, context or relationship to others. In contrast, externalised norms abstract away from these details. In other words, normative criticism that is not externalised carries uncertainties that expressions about externalised norms do not, and it is this clarity and precision which is advantageous even when interacting with in-group members. When expressing that stealing is wrong, one takes it to be context- and authority-independent unless

otherwise specified. It therefore more directly advertises the sender's expectations regarding the behaviours of her interactive partners even within groups with shared cultural norms.[87]

A final advantage of moralised commitments is that we take moralised language to be especially meaningful for predicting an agent's future behaviour, increasing $P_C(R1)$. For example, one will be surer of the sender's future course of action should she say, "it would be wrong for me to not help you with your fig-gathering" than if she said, "I will help you with fig-gathering". Indeed, Peters and Kashima (2013) as well as Jordan and colleagues (2017) find evidence that those who either directly assert their moral values or condemn the immoral behaviour of another were judged to be more moral themselves, or less likely to commit the stated transgressions. Jordan and colleagues (2017) presented participants in the study with a series of vignettes in which subjects were asked to imagine that they belonged to a social group in which a particular moral transgression was possible, for example, a track team whose members could use a forbidden performance-enhancing drug. To test whether moral gossip was an indicator of future behaviour, the participants were presented with a vignette of an agent condemning drug use and a vignette of them not condemning drug use. They then rated the targets on their likelihood of committing the relevant transgression, trustworthiness in this context or new contexts, and likeability. It was found that subjects evaluated targets more positively when they engaged in condemnation when they did not. In one of the conditions, they were also told that the agent had never used such drugs. The

[87] Often, signalling one's attitudes concerning disgust will also take on an externalised character. The emotion of disgust is experienced as a strong feeling of revulsion and causes a motivation to withdraw from or avoid the stimulus (Rozin et al. 2000). Disgust can be physical, sexual or moral (Massar 2021). It has been hypothesised that this emotion serves to facilitate social distancing and avoidance of those believed to commit the disgust-eliciting transgression (Tybur et al. 2009). If this is so, disgust signalling also serves to signal one's future behaviour and reveal one's expectations concerning other's actions. As mentioned earlier, our concern is not only with moral norms but any norms we take to be externalised and norms concerning eating rotten meat or incest may similarly be externalised.

positive evaluation was less strong in the condition where they had prior information about the agent, supporting the suggestion that the *function* of the condemnation was to signal moral behaviour (since in the good-information condition, condemnation is no longer necessary).

This increased judgment of the speaker's morality is likely related to the proximate motivations of externalisation. Since externalised norms are typically moral, they will often involve motivationally salient proximate mechanisms such as guilt, shame, and pride which incentivise an agent to follow through on their commitment. Indeed, as we transition to larger and more complex societies, fear of ostracism will become a less effective means of ensuring that an agent follows through on her commitment. This is because, in order for reputation to incentivise an agent to follow through on her commitment, it must be the case that news of the agent's defection will reach others with whom she is likely to interact. With growing social networks, this likelihood shrinks. So we need some other means of providing this incentive. Externalisation provides just such a proximate motivation. Indeed, even if the ultimate cause of cooperative behaviour is exclusion from future benefits, the attainment of future benefits might be best secured by genuine moral dispositions – in this case, viewing a norm as externally imposed. This is not to say fear of ostracism does not play a role.[88] Yet, if the emotions associated with norm-externalisation incentivise the sender to follow through on her commitment in the absence of reputational effects, moralised commitment will promote cooperation more reliably. As such, the receiver will form more robust expectations of the sender's future cooperation.

[88] Note, however, that our attitudes toward ostracism may not have been the initial proximate mechanism underlying an agent's motivation to follow through on a commitment, since these attitudes likely coevolved with the deceptive use of implicit commitment signals in order to limit susceptibility to deception. When moralised commitment first emerged, the proximate mechanism for commitment behaviour likely turned on the unique phenomenology of externalised norms (Stanford 2018a; 2018b).

The ability to make moralised commitments affords more opportunities for interaction in which cooperation or defection is meaningful for determining a partner's reliability. The more commitments made, the more data we have on whether a partner can be trusted since we have more instances to see whether they follow through. The more data we have on whether a partner can be trusted, the better we secure correlated interaction among cooperators in a population. Greater correlated interaction reduces the likelihood that cooperators are invaded by defectors. Thus, the ability to make these commitments ultimately allows cooperative tendencies to spread. This is particularly important given the extension of the cooperative enterprises we have undertaken over the course of our evolutionary history, and the new risks of exploitation this entails.

In sum, at some point in our evolutionary history, we began to consider some norms as universally applicable or externalised (Stanford 2018a). The advent of norm externalisation allowed us to signal commitment to cooperation via signalling our attitudes on behaviours taken to be universally obligatory, either directly, or indirectly through gossip or story-telling. As with our previous forms of commitment, this both introduces a fitness cost to sender defection and enables the identification of other cooperators. Moralised commitments extend the scope and power of commitment. Expressions of attitudes toward externalised norms make implicit commitment independent of the cultural background and role of the sender of the commitment signal, meaning these commitments are easier to make and to identify. Furthermore, by making a moralised commitment, the sender simultaneously advertises her own trustworthiness and reveals to the receiver that she is willing to exclude others for not acting in like manner. She thereby secures better correlated interaction by

protecting herself from interactive partners who would act otherwise. Finally, moralised commitment reliably changes receiver expectations of the sender's future actions and this is likely linked to the unique phenomenology of externalisation and associated proximate motivations for following through on one's commitment. These advantages make commitment more effective and thereby secure better correlated interaction, extending the scope of successful cooperative interaction.

4.5 Securing more cooperation

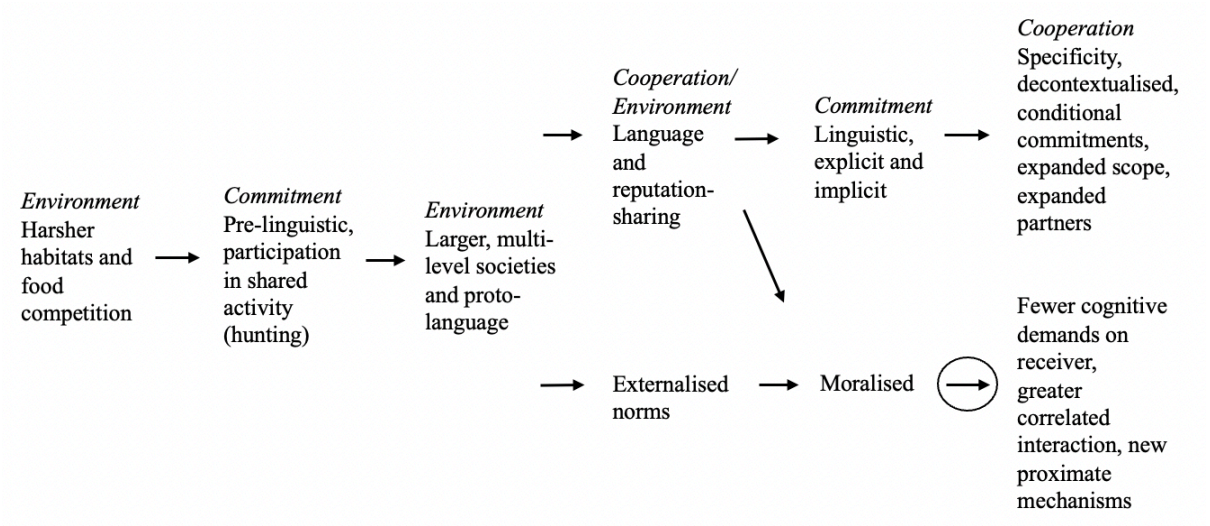

Figure 16: Expanded cooperation as a result of moralised commitment.

We have a second coevolutionary link: the cognitive capacities which coevolved with our early shared activities set the stage for the emergence of moral cognition, both in further developing our perspective-taking capacities and strengthening our affective mechanisms. Alongside the larger and more complex societies we then faced, this led to the emergence of norm externalisation as a means of correlating interaction. This had important consequences for our practices of

commitment. Commitments could be made by expressing one's attitudes toward externalised norms, both directly, and indirectly via gossiping about another's norm-violation or adherence or via story-telling concerning externalised norms. The feature of the utterance that makes this a commitment is that it simultaneously changes receiver expectations of the sender's future course of action and changes the sender's relative payoffs since she is subject to potential exclusion from the benefits of current and future interaction should she act otherwise. Deception will be limited by the same mechanisms we saw in the previous chapter on linguistic commitment.

Importantly, this kind of moralised commitment has additional fitness advantages and better secures correlated interaction than our previous forms of commitment. First, it allows us to abstract away from details about the sender's cultural background and social role when attributing commitments. This makes it both easier for the sender to signal commitment and easier for the receiver to identify such commitments. Second, moralised commitment is not only used to advertise the sender's future behaviour but it also reveals that they hold others to the same norms, since externalised norms are taken to apply to all, irrespective of other features of the agent. This means that, as long as the receiver of the commitment signal takes the norm the sender is espousing to be one that the sender externalises, the receiver will understand herself to be open to potential exclusion if she does not act in a similar manner. As such, she is deterred from defection herself. This secures better correlated interaction on the part of the sender of the commitment signal since she makes her expectations of others clear at the same time as she changes others' expectations of her own actions. Note that this will apply within the group as well as outside the group. Cultural norms do not achieve the same change in expectations since the judgment the sender passes on the norm-adherence or -violation may depend on the social role of the actor involved. Finally, moral

language and judgments carry with them powerful motivational forces which may more reliably communicate the sender's future course of action than non-moral language. These fitness advantages help to explain the emergence of moralised commitment as a means of securing correlated interaction in a world where our cooperative interactions extend across multi-level societies and interaction with new partners is rife. We thus expand the range of successful cooperative interaction with the advent of moralised commitment.

In the next chapter, we see how, in the context of increasing cultural development enabled by previous forms of cooperation, we were able to create institutions which reflected and codified commitment practices such as legal institutions and religious bodies. These provided a new means of enforcement for commitments – not only could one's reputation be on the line, but also one's physical and financial security. The cost of reneging on one's commitment has thus responded with increased voracity in response to an ever more interactive world. That is, I show how our forms of commitment once more coevolved with cooperation.

Chapter 5: Institutionalised commitment

We have, already, the ability to commit to cooperation by engaging in shared activity, by linguistic promising, both explicit and implicit, and by making moralised commitments with reference to externalised norms. Commitments can be made among strangers within one's group and concerning the granular details of interaction. The cooperative environment we now find ourselves in is one of prolific cooperation in large groups. In this chapter, I argue that our trajectory to modern human prosociality, which is widespread and spontaneous, crucially involved an increase in the clarity and specificity of commitments, and a lowering of the costs of enforcement of commitment. A particularly salient way that we have achieved this is through the institutionalisation[89] of commitment. By institutionalised commitment, I mean that these commitments are enforced by third-party punishment conducted in an organised manner, it is common knowledge that transgressive behaviours will be punished and, typically, the punishment is carried out by a body which undertakes a policing role in the community and is funded by its members. Our previous forms of commitment were backed by one's reputation – this is what caused the change in sender payoffs at the level of ultimate causation. However, there will also be externally enforced commitments where the cost of punishment has been outsourced to third parties, for example, legal bodies, religious bodies, or cultural group policing. Note, third-party

[89] Currie et al. (2016) write "we conceive of institutions as human-generated regulators of social interaction and adopt a working definition of institutions as systems of interrelated rules which prescribe particular roles and regulate social relations. Examples of institutions would be marriage, descent and inheritance systems, codified legal systems, parliaments, and banking… Formal institutions are equated with written rules and enforced by a disinterested third party" (Currie et al. 2016: 200). Of course, this is a broad notion of institutionalisation and for our purposes, the definition will be narrower.

enforcement is a necessary but not sufficient part of formal institutions – there is much else that makes an institution formal, which we will get into detail on later.[90]

Importantly, institutionalised commitment also *coevolved* with previous forms of cooperation. In particular, a common language extended the broadcasting of explicit and implicit commitments in a public setting among strangers. Onlookers then had an opportunity to enforce these commitments, perhaps to bolster their own reputation among their subgroup. The rise of more inegalitarian societies and greater inter-group competition during and following the Neolithic revolution would have intensified the demands of cooperation alongside providing more powerful hierarchies to enforce commitment. At the same time, institutions were becoming more formalised – written, impersonal and organisational. By the time we had the modern policing bodies of governments and courts, the threat of third-party punishment was a clear constitutive part of an institutionalised commitment and this had profound impacts for cooperation. So we add the following link to our coevolutionary story (Figure 17). In this chapter, I will discuss the emergence of third-party punishment of commitments, increasing institutionalisation, the new forms of commitment this enables, and the effect this had on cooperation.

[90] Indeed, indirect reciprocity is a form of third-party punishment which is not formally institutionalised. Yet we will see that institutionalisation allows more damaging forms of punishment at low cost.

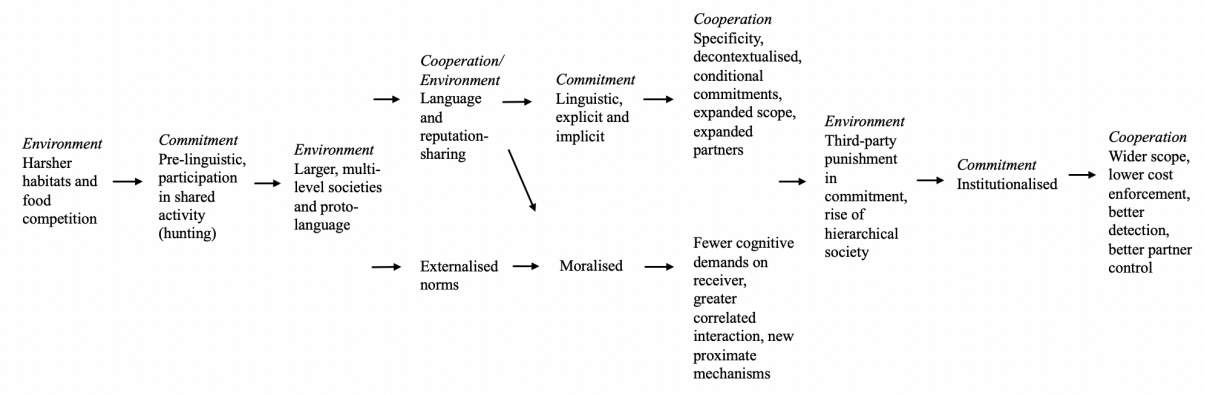

Figure 17: The coevolution of commitment and cooperation.

5.1 Third-party punishment of commitments and hierarchical society

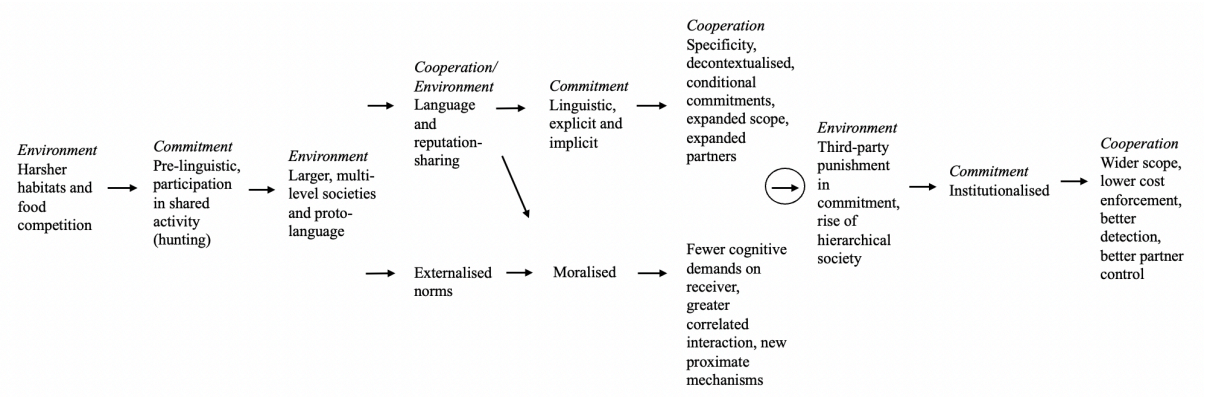

Figure 18: Third-party punishment of commitments and the rise of hierarchical societies.

Once we had commitments via shared activity and linguistic commitments, agents were able to engage in correlated interaction on precise matters and were able to make commitments with strangers, across time, and with multiple means of ensuring their credibility. Agents could also solve collective action problems which would likely not have been possible before the advent of linguistically communicated commitment, and their actions were guided by externalised norms

applicable to individuals across groups. In this section, I argue that these changes laid the groundwork for the proliferation of third-party punishment in commitment. I suggest third-party punishment of commitments likely first evolved in cases of group enforcement of publicly observable commitments. I then discuss how the demands of agricultural life in the Neolithic created selective pressure for a more powerful enforcement mechanism than reputation and thus contributed to the proliferation of third-party punishment of commitment. I also address how the Neolithic transition was aided by our previous forms of commitment and cooperation. With the rise of inegalitarianism shortly afterwards, and the simultaneous formalisation of institutions, third-party punishment became a powerful and organised policing mechanism to underwrite commitment.

The purpose of this section is not to explain the evolution of third-party punishment simpliciter. The target of the current discussion is to understand when and why such punishment was directed towards those who reneged on commitments, and why it was institutionalised. Explaining the evolution of third-party punishment simpliciter is a much larger task than can be accomplished in one chapter.[91] What matters most for our current discussion is that people *do* punish those who hurt others. Fehr and Fischbacher (2004) studied a third-party punishment game with three players. The game played between players A and B is a dictator game where player A can assign any portion of their endowment to player B. Player C has the option of punishing player A for their assignment, where punishment costs player C one point for each three points deducted from player A (converted into monetary outcomes at the end of the game). Of course, a self-interested player would never punish player A. However, this is not what we see in practice. In the experiment,

[91] For a discussion on this, see Boyd et al. (2005).

approximately 60 per cent of players would punish whenever player A transferred less than half of her endowment to player B, and this punishment would scale with how unfair her transfer was.[92] There is also evidence that third-party punishment is seen cross-culturally (Henrich et al. 2006).

One might think that third-party punishment requires that the agent who is doing the punishing is a disinterested party. On this reading, group punishment of violated commitments might not be an instance of third-party punishment since each agent in the group has a long-run interest in the behaviour of other members of the group since they may interact with them in the future. They thus have an interest in encouraging prosocial behaviour from an agent who has not directly harmed them. Even in larger societies, where we may not interact with the same individual, we have an interest in deterring the possibility of anti-social behaviour so would pay an insurance premium to have that incentive protected. If this is so, explaining group punishment is not a difficult problem.[93] Indeed, Zhang and colleagues (2014) show how pool-punishment (where individuals contributed to a common pool in advance used to punish others) was chosen over peer-punishment in experimental set-ups, as long as punishment is also imposed on second-order free-riders. Schoenmakers and colleagues (2014) extend this analysis via modelling work to show that the punishment of second-order free riders is not strictly necessary to ensure the establishment of sanctioning institutions as long as these institutions are publicly visible, acting as a costly signal which deters defection. Consider, for example, a police station in a central location. However, whether we deem these sanctioning institutions as truly enacting disinterested punishment is not central to my account. What matters is that others in the group enact punishment, and I seek to

[92] Fehr and Fischbacher (2004) also study third-party punishment in a Prisoner's Dilemma game and find almost half of the third parties punished if one of the players defected. Only approximately 20 per cent punished if both players defected, suggesting this was a less severe norm violation.
[93] Though the second-order problem of altruism exists.

explain how this phenomenon could have emerged in our past. For this reason, I will use the term "third-party punishment" to refer to punishment enacted by those other than the individual who was harmed, regardless of whether this punishment is ultimately in the punisher's self-interest.

To begin our discussion of the emergence of third-party punishment of commitments, we will first note that one way of ensuring credible commitments discussed in Chapter 3 is making commitments publicly. If commitments are made publicly with strangers, onlookers can enforce the commitment by social exclusion (as long as subsequent cooperation or defection is also observable). Social exclusion is a low-cost means of enforcement and makes the sender's commitment more credible since she faces a cost to reneging. Indeed, hunting is an example of a publicly observable commitment, but the agents who enforce the commitment are those same agents which are affected by reneging, so there is no difficulty in explaining the incentive to punish here.[94] However, in the case of public linguistic commitments, the onlookers are not involved in the commitment game that the dyad is playing.

I suggest these onlookers are engaged in a different strategic interaction. The game that the onlookers are playing may involve costly signalling of their virtues through enacting costly punishment. Jordan and colleagues (2016) develop a model in which they show that enacting costly third-party punishment can serve as just such a signal of trustworthiness to others. They empirically validate this model by showing that third-party punishers are trusted more in experimental setups where they punish a "helper" for an insufficient contribution to a recipient in a Dictator Game. This increased trust is demonstrated by a subsequent Trust game played with the

[94] At least when social exclusion is low-cost. See Chapter 2 for a discussion of costly punishment in the context of public goods.

punisher, where others are more likely to trust them after they have engaged in third-party punishment. It was found that agents sent 16 percentage points more of their endowment to third-party punishers than non-punishers. The punishers themselves also behave in a more trustworthy manner in the subsequent game, showing that this increased trust in them is warranted. It was found that those who punish return 8 percentage points more of the entrusted endowment to the original agent than those who do not punish. If punishment can act as a costly signalling of one's trustworthiness within one's subgroup, this may be and explanation of why third parties enforce commitments in dyads even when their interests are not directly harmed by the foregone commitment.

We are yet a long way from modern, institutionalised punishment. For our intermediate steps, it is important to note that explicit and implicit commitment almost certainly preceded the Neolithic revolution approximately 12 kya and aided in the change itself.[95] This had important impacts for the growth of third-party punishment and institutionalisation. The Neolithic revolution is characterised by a transition to less mobile foraging, more permanent settlements, and an economy based on domesticated livestock and cereals. There was also an expansion in the production of artisan goods and regular trade, making commitment even more important (Sterelny 2019). Crucially, this revolution was, in part, made possible by our previous forms of commitment and cooperation.

[95] Powers et al. (2016), for example, estimate the arrival of language at 500 kya (though later estimates place it at 100 kya) and the Neolithic revolution at 12kya years ago, as does Stuart-Fox (2022). See also Watkins (2010) for detailed periodisation.

We saw in Chapter 2 that hunting, among other shared activites, allowed us to live in larger groups, with greater means of provisioning, a division of labour and extended life spans, all of which were precursors to agricultural life. I also argue in Chapter 3 that explicit commitment facilitates this transition to a more settled, agricultural life since it is only with the ability to communicate about specific divisions of labour through language that we are able to make such gains. There are a multitude of tasks to do in a collective action problem such as farming – the clearing of land, planting of crops, harvesting and storage were not typically individual activities (Sterelny 2019). When this is so, linguistic communication becomes an important way of ensuring cooperation. Not only this, but increasing complexity in tasks demands increasing plasticity of response and therefore more flexible means of communication. Linguistic communication about role expectations allows us to respond plasticly to contingencies we may not have faced before, where conventions based on shared activity could not have. Furthermore, moralised commitment increased the scope of our cooperation, providing selective benefits to groups who were able to employ this as a means of correlated interaction, and extending cooperation more safely to strangers by introducing a new means by which commitments are incentivised – the phenomenology of externalised norms. With cooperation possible among strangers, we can solve collective action problems with novel partners. As such, our previous practices of commitment and cooperation contributed to a change in our cooperative environment which, as we will see, selected for an increasingly effective form of commitment.[96]

[96] Other factors which contributed to the Neolithic revolution include population growth, climate change, the availability of suitable wild species, territorial control and defence. See also Stuart-Fox (2022) for more factors in the Neolithic transitions.

The way in which the Neolithic revolution changed our cooperative landscape created selective pressure for the strengthening of our systems of third-party punishment. Sterelny notes that cooperation would be more stressed in Neolithic life. He writes, "cooperation would still be essential; probably more essential, given the demands of collective action. But it was much more stressed: by the increase in community size; by the decline in intimacy; by the burdensome nature of collective labor; by the time depth of return on investment and the risks imposed by delayed return" (Sterelny 2019: 8). In particular, a decline in intimacy and growth in the community size would have diminished the effectiveness of reputation as an enforcement mechanism for commitments. So the demands of collective action in Neolithic life alongside the increased difficulty of effective correlated interaction increases the importance of developing a new enforcement mechanism to ensure commitments are followed through. If the practice of third-party punishment is already at play in these societies (in the form of social ostracism for publicly violated commitments), this affords us with an enforcement mechanism which can be used to secure commitment in the more complicated collective action problems we then faced.

The emergence of more segmented and hierarchically organised societies would have strengthened third-party punishment further. Unsegmented societies are those in which social organisation is usually horizontal, across members of the band, whereas segmented societies involve vertical organisation. Individuals are members of families, lineages and clans. Segmented societies will also involve trade and cooperation across bands. With cooperative interaction taking place at a larger scale and without recourse to reputation within one's group as the means of enforcement for commitments, there is again selective pressure to develop another means of enforcement. Importantly, with a segmented, hierarchical structure, there is also scope for more powerful third-

party enforcement since elites are more likely to have the wealth and resources to enact punishment beyond social exclusion – for example, fines or imprisonment. Not only this, but it is more likely that others will unite in support if there are clearly defined leadership roles for the initiation of punishment and where signalling one's membership to others in the group is important. Indeed, in clan-based communities, one's social capital largely comes from alliances with clan-mates rather than reciprocal partners, meaning clan-mates are invested in supporting one another.

The rise of transegalitarianism likely occurred after the demise of the Neolithic communal structures around 10 kya. One of the major contributing factors was the emergence of a clan structure with a strong corporate identity, based on a purported genealogical connection and entrenched by intense initiation rites or collective activities (Sterelny 2021c). Individuals get social support mainly within the clan. Once we had storage and farming, elites could use this clan membership to garner social and material support. Indeed, it can be argued that farming itself contributes to inequality since it establishes private property, and storage undermines the need for sharing as a means of smoothing risk (Sterelny 2021c). A second major factor in the transition to transegalitarian society was the development of quasi-elites who have expertise or control information concerning subsistence skills (such as navigation, plant identification, or artisan skills) (Henrich 2015). This control of information can lead to deference, particularly if information transfer takes the form of emotionally-charged ritual or religious narratives.[97]

[97] Transegalitarian communities also typically emerge in contexts of intercommunity violence, again linked to the rise of sedentism, since this allows us to accumulate material wealth in a way that foraging could not (Sterelny 2021c). A modern ethnographic equivalent of the transegalitarian society is the Big Men society of Melanesia (Sahlins 1963). Big Man status is not a political title but is an acknowledged standing in personal relations, which attracts the loyalty of others. Sahlin (1963) notes that Big Men typically get what they want through public verbal suasion.

This transition to a transegalitarian world where elites were common is important as Ozono and colleagues (2016) show how, in Public Goods games, leaders in a group can stabilise cooperation by punishing both non-cooperators and non-supporters (those who do not punish other non-cooperators) and that leaders in fact benefit when this is the case, favouring the evolution of punishment and second-order punishment. Supporters will be incentivised to punish renegers of commitment if they will incur a cost for non-punishment and if their punishment signals trustworthiness to others. It is also possible the cost of punishment is lower for group leaders, since they are most likely to get support and will not therefore be the sole target of potential retaliation. As such, onlooker punishment of commitment in this type of society will become more organised.

So we have seen how third-party punishment may scale up further in the Neolithic revolution and beyond, and how early forms of commitment and cooperation contributed to this development. This is not a step-wise transition but rather a gradual coevolution of our cooperative lifestyle and our punishment practices. Indeed, chiefdoms were perhaps a transitionary step between villager societies and larger political units, and these involved an intermediate increase in the power of elites to enforce punishment. Moreover, as Stuart-Fox (2022) writes, "sedentism is not an all-or-nothing condition" (Stuart-Fox 2022: 22). Most hunter-gatherers in the Levant had bases with seasonal food resources to which they returned. In short, the increasing cooperative demands of sedentary life which created further pressure for third-party punishment of commitment likely also incrementally selected for increased organisation, frequency and strength of punishment as our cooperative enterprises and hierarchical structures developed.

We have shown how (i) third-party punishment of commitment may have arisen among agents interested in signalling their trustworthiness to others within their subgroup, (ii) how the transition to the Neolithic revolution would have made an effective enforcement mechanism such as third-party punishment even more important given the complexity of interactions, (iii) that this revolution was in part made possible by previous practices of commitment and cooperation, and (iv) that the rise of transegalitarian societies after the Neolithic revolution would have provided further resources for enacting costly punishment with the uniting power of group leaders. However, we have not yet addressed the *institutionalisation* of such punishment.

First, it is important to note that there is a difference between informal and formal institutions. This difference is on a continuum and spans multiple dimensions. Generally, formal institutions are characterised by constitutions, contracts, and forms of government (Kaufmann et al. 2018), whilst informal institutions will include "traditions, customs, moral values, religious beliefs, and all other norms of behavior that have passed the test of time" (Pejovich 1999: 166). More precisely, while informal institutions typically govern personal exchange, are unwritten, non-contractual, are premised on shared expectations and are enforced privately, formal institutions govern impersonal exchange relations (for example, trade), involve written, contractual rules of law, are premised on organisational goals (for example, societal order), and involve third-party enforcement (Hyden 2006). Of course, institutions can be more or less formal depending on whether they have some or all of these features and the sophistication of these features. For example, Australian aboriginals have complex kinship systems accompanied by explicit norms. This institution will be more formal than one where the norms of interaction are not explicit, even though it does not involve impersonal exchange relations. So an institution may be more or less formal in one of the relevant aspects of

formality without being so in another. Furthermore, though there may be instances of formal institutions which do not involve third-party punishment, all that will be relevant to our current discussion do.

Like the emergence of third-party punishment in commitment, institutions are also the product of evolutionary processes and the transition between formal and informal institutions is not step-wise but gradual. One might already note that informal institutions would have been present in early hunter-gatherer societies in the form of obligations in kinship relations and food-sharing rules. These informal institutions would have aided in the transition to the Neolithic revolution and the development of more formal institutions by creating shared expectations of cooperation on which more complex interaction could be based. However, the more complex the interaction, the greater the need for formalisation to safeguard cooperation, since institutions make mutual expectations explicit, reducing ambiguity and transaction costs.

Turchin and colleagues (2013) present a model of the rise of formal institutions as a result of inter-group conflict. The key idea is that institutions serve to make severe punishment cheap enough to incentivise high-cost cooperation in military activity. In conditions of warfare, the costs of maintaining an ultrasocial institution are outweighed by the benefit that institution provides at the group level. The authors speculate that increased warfare could be traced back to war between settled agriculturalists and nomadic pastoralists in the steppe regions of Eurasia, and the development of horse-based military technologies such as chariots and cavalry. Regardless of whether this is the correct story of the transition, it is very likely that institutions were becoming increasingly formal just as third-party punishment was becoming even more important for securing

cooperation – war represents one paradigm example of this relationship. Indeed, some formal institutions are directly selected for by the advent of large-scale agriculture in the Neolithic revolution. Notably, private property rights and hierarchical rules concerning the depletion of common-pool resources (Powers et al. 2016).

Formal institutions offer a number of fitness advantages. These fitness advantages would be realised by cultural selection between groups.[98] First, where there are multiple equilibria in a cooperative interaction, institutions can aid in equilibrium selection, leading us to the outcome with the highest payoff (Powers et al. 2016). Second, where individuals are subject to time-discounting and may choose irrationally as a result, institutions can set up incentive structures that promote optimal long-run choices. Third, where information about others in one's group is not readily available given growth of the group size, institutions or institutional membership may provide such information. Fourth, in relation to agriculture in particular, groups with institutional rules regulating irrigation have been demonstrated to successfully solve collective action problems (Powers & Lehmann 2013). Institutional rules also regulated trade of staple goods during the Neolithic (Oka & Kusimba 2008). These rules could have contributed to the development of larger polities as it pays to interact with others who play by the same institutional rules (Powers et al. 2006). Finally, and most importantly, the development of formal institutions comes to shape the role of third-party punishment.

[98] We would expect cultural group selection to be the evolutionary explanation at play when there are mechanisms which sustain between-group variation. In particular, where there is conformist transmission (the psychological bias to imitate high frequency behaviours); prestige-based transmission (the tendency to copy the successful); self-similarity transmission (the tendency to copy the similar); normative conformity (the act of choosing one's behaviour simply to appear to match the majority) and punishment of non-conformists. There must also be mechanisms that affect a group's proliferation. These may be demographic swamping (faster reproduction of one group in the region); intergroup competition; or prestige-biased group selection (when individuals have the opportunity to copy others from nearby groups, they imitate those that are more successful). For more details on the conditions under which cultural group selection is applicable, see Henrich (2004).

With the development of community bodies which upheld norms concerning interaction, third-party punishment has scope to become *institutionalised.* By this, I mean that the punishment is conducted in an organised manner, it is common knowledge that transgressive behaviours will be punished and, typically, the punishment is carried out by a body which undertakes a policing role in the community and is funded by its members. To say that a commitment is institutionalised, then, is to say that it is made under conditions of institutionalised third-party punishment. Note that we do not need to have modern legal bodies in order to consider third-party punishment institutionalised. Minimally, we require that there are cultural norms present in the group, that these norms support enforcement of commitments, and that costs for enacting these norms fall on other members of the community. So there is a continuum of institutionalised third-party punishment. Of course, as we develop more precise structures regulating interaction, we will see third-party punishment become more powerful. Eventually, we arrive at modern governments and courts which can enact third-party punishment on behalf of agents where commitments are not upheld.[99] In the next section, I will present examples of minimally institutionalised and more formally institutionalised commitments.

So linguistic commitment made in public settings selected for third-party onlookers, playing their own reputation game, to be involved in the enforcement of commitment. Our previous forms of commitment also provided the cooperative environment which made possible the Neolithic revolution. This made the proliferation of third-party punishment in commitment more important.

[99] One might ask: why would an institution do the punishing? If cost is distributed widely among a population – and if, in fact, such costs are required to be a member of the community and to partake in public goods, agents will pay them since to not pay is to be excluded from the safety that the group and all its institutions have to offer.

In particular, hunting and other shared activities allowed growth of the group size and a division of labour. Explicit promising allowed us to coordinate on complex collective action problems which demand plasticity of response. Moralised commitment increased the scope of cooperation, extending cooperation more safely to strangers with a new means of ensuring the agent follows through on her commitment. With the advent of a more settled, agricultural life, collective action problems increased in frequency and complexity, yet interaction with strangers was common and reputational means of enforcing commitments became stressed. This created selective pressure for the strengthening of third-party punishment to act as a means of enforcing commitments. Third-party punishment would then become most stable in transegalitarian societies with clear leaders. With the rise of elites, third-party punishment became an even more powerful enforcement mechanism. At the same time, institutions were becoming more formalised. This meant that third-party punishment was more organised, more precise, and lower cost. With this change in our cooperative environment in mind, we can now talk more precisely about how institutionalised third-party punishment would have affected our practices of commitment and cooperation.

5.2 Institutionalised commitment

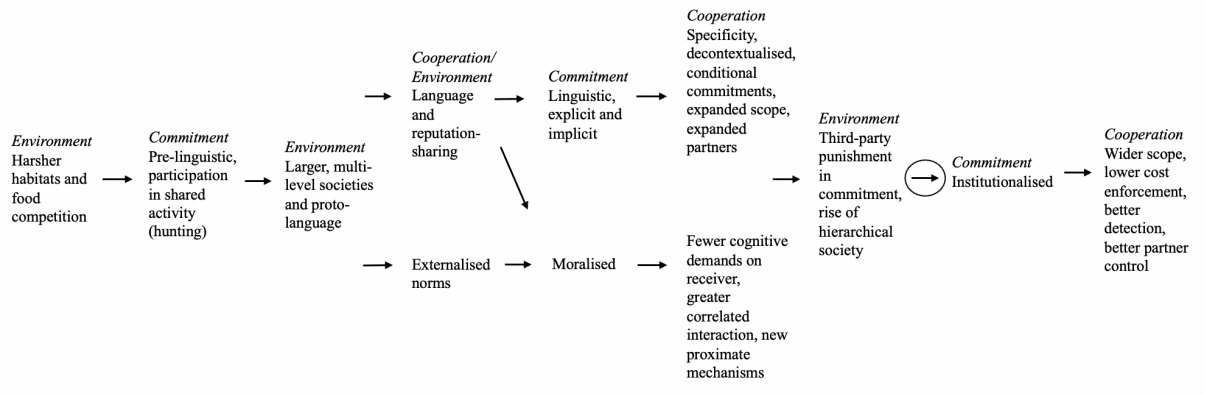

Figure 19: The emergence of institutionalised commitment.

Institutionalised commitments can take many forms but the unifying feature of these forms of commitment is that the change in sender payoffs are not (only) a result of a tarnished reputation. Rather, the cost of reneging on one's commitment can include financial costs, imprisonment, or other punishments. Importantly, institutionalised commitments involve costs which are imposed from agents *other* than the receiver of the commitment signal. These differences are important, since the greater the cost of reneging on one's commitment, the more credible the commitment is to the receiver. While social exclusion was a low-cost form of punishment, other financially or physically punitive measures can increase the potency of potential punishment while simultaneously reducing costs for the receiver. Sometimes these costs are imposed coercively and sometimes these costs are willingly incurred by the sender themselves in the form of a commitment.

In the previous section, we discussed how some institutions may not have all features of formalisation. Here, I consider one where the condition of a policing body is not met. Mathew and Boyd (2011) illustrate how community-imposed sanctions for cowardice and desertions in war maintain cooperation in the Turkana community, a politically uncentralised, egalitarian, nomadic pastoral society in East Africa (Mathew & Boyd 2011). Here, there is no designated body which carries out a policing role, yet punishment is minimally institutionalised in that punishment is conducted in an organised manner in conditions of common knowledge, and the cost for enforcing these norms fall on other members of the community. At least one person was sanctioned in 47 per cent of raids where desertion was reported and in 67 per cent of raids where cowardice was reported. Desertion and cowardice are forms of reneging on the commitment to participate fully in the raid. Acccusations of defection require consensus from the group and punishment is carried out by third parties – members of the violators age-group – even if they did not participate in the raid and did not experience the consequences of the violation. Serious cases involve not only verbal sanctions but severe corporal punishment, where the defector is tied to a tree and beaten by his age-mates.

Increasing in formality, but still related to war, we see the institutionalised commitment of Roman military oaths. The *sacramentum militare* (also as *militum* or *militiae*) was the oath taken by soldiers in pledging their loyalty to the emperor in the Republic era onwards. The oath took the form: *Iurant autem milites omnia se strenue facturos quae praeceperit imperator, numquam deserturos militiam nec mortem recusaturos pro Romana republica*! ("But the soldiers swear that they shall faithfully execute all that the Emperor commands, that they shall never desert the service, and that they shall not seek to avoid death for the Roman republic!"). If the soldier reneged

on this commitment, he was subject to harsh penalties, such as execution or corporal punishment enacted by a policing body.

An example of a modern institutionalised commitment is a legal commitment such as a lease agreement. Legal commitments are primarily contractual rather than subjective commitments – ones where the mental state of the agent is (relatively) unimportant (Frank 2003). Contractual commitments are externally enforced. To elucidate, Sandy (the sender of the commitment signal) may enter into a lease agreement with Betty (the receiver). Here, Sandy is the tenant and Betty is the landlord. After making this commitment, Sandy's payoffs have changed for cooperation, but the punishment for defection need not be enacted by Betty – it can be enforced by an institution. In particular, the National Residential Landlord Association or court system. Betty's expectations of Sandy's intended action have changed in the way we would expect of a commitment – she expects rent – but, unlike our previous forms of commitment, Betty is not the only one to whom Sandy is liable. Using our earlier definition, *a commitment is a pre-play signal in a strategic interaction taken at time t that increases the sender's relative payoff for carrying through option X at time t+n, and increases the receiver's probability of the sender carrying through option X.* Here, the pre-play signal is the undertaking of the legal contract and option *X* is paying rent. Time *t+n* indicates when the rent is due. In this particular example, the lease agreement may be a mutual commitment, with Betty offering to provide basic repair services for Sandy. The third-party punisher in this case is the Tenant's Union. As in previous chapters, the payoff change at the proximate level is captured in the extensive form game from Chapter 1, and in the evolutionary game at the level of ultimate causation.

Contracts can be unilateral or bilateral (involving promises by only one player or both). The offeree is typically entitled to financial compensation for some service performed and if the offeror refuses to pay, a court will decide whether there was a breach of contract based on the clarity of the terms. An example of a unilateral contract is a reward contract, where someone posts a reward for their lost pet. In bilateral contracts, the receiver of a commitment signal is afforded additional means of enforcement. She may invoke the law to enact punishment on the sender and this punishment may take the form of *partner control,* rather than partner choice, the latter of which is premised on selecting agents with a good reputation. For example, in legal agreements which mandate receipt of an item in exchange for some service, if the agreement is violated, the law can intervene to repossess the item and give it to the receiver. If this is so, the enforcement takes the form not of a penalty but of reparations such that the commitment is forcibly followed through. This differs slightly from our previous forms of commitment in that partner selection based on commitment does not hold as central a role, since partner control can be ensured by a third-party.[100]

Contract law institutionalises commitments by providing codified third-party enforcement, lending weight to the credibility of one's commitment and making more explicit what counts as cooperation. Fried (1981) in particular points to the role of the law in making possible mutual benefit in cases of reciprocal exchange over time rather than, as Sterelny (2021a) puts it, immediate return mutualism. As noted in Chapters 2 and 3, it is more difficult to explain cooperation in this

[100] However, it is important to note that one could describe partner choice and partner control as two sides of the same coin. Indeed, Wubs et al. (2016) argue that the three main partner control mechanisms are: "(i) to switch from cooperation to defection when being defected ('positive reciprocity'), (ii) to actively reduce the payoff of a defecting partner ('punishment'), or (iii) to stop interacting and switch partner ('partner switching')" (Wubs et al. 2016: 1). All three mechanisms seem to be at play in the commitments we have talked about in previous chapters. We have focused on partner switching and labelled this partner choice. The difference is that tort law allows reparations equivalent to the commitment to be forcibly made, in a way that seems disanalogous to the commitment mechanisms we encountered previously.

case as it requires trust in another individual over time and is more susceptible to cheating. Fried writes "we need a device to permit a trade over time: to allow me to do A for you when you need it, in the confident belief that you will do B for me when I need it. Your commitment puts your further performance into my hands in the present… in order to accomplish this all we need is a conventional device which we both invoke" (Fried 1981: 13-4). This conventional device takes the form of institutionalised commitment.

5.2 The advantages of institutionalised commitment

What fitness benefit do these institutionalised commitments afford us at the level of sender and receiver? Most importantly, for the receiver, the value of *l* is lowered relative to the value of *c*. Exclusion was an already low-cost form of punishment but enacting *further* punishment in terms of financial penalty, imprisonment or restitution would be very costly for an individual agent to enforce. This is now, however, enacted by a policing body. Contributions to the upholding of such a third-party policing mechanism are small compared to the difficulty an individual agent may go through to acquire her promised goods. Importantly, *c* increases for the sender, disincentivising acting in contradiction to one's commitment. The cost now includes not only reputational consequences, but physical and financial ones, too.

Furthermore, as mentioned earlier, an agent can also be forcibly made to follow through on their commitment, rather than simply punished for not doing so. For example, in a sale agreement. Alternatively, if the good is not provided, we can be assured that appropriate compensation is received for a reneged commitment via debit or credit card chargeback. In such cases, agents do not need extensive information about the reputation of a new partner in order to consider it

worthwhile to interact with her. As long as amends can be forcibly made, an agent's prosociality is institutionally protected. One does not need to trust that the seller is reliable since one is safeguarded. However, this will not be true of all institutionalised commitments. For example, taking a landlord to court for not providing responsible care of one's property is likely to be a costly and time-consuming process. In these cases, access to reputation may serve to bolster trust in new partners. Nonetheless, the complexity of this interaction is better safeguarded by means of institutionalised commitment than it is without.

Though not directly related to the introduction of third-party enforcement, institutionalised commitments may also offer other fitness advantages. For example, in legal commitments, the terms of a promise are typically made clearer than in our previous commitment mechanisms, usually in a written agreement that can be referred to later. Note that this is not always the case. A non-legal commitment can also be written and a legal contract could be oral or poorly defined. This is only a claim about what we typically see today. If legal commitments are more precise, this translates into increased $P_C(R1)$ conditional on the signal, since the signal is more reliable concerning what counts as cooperation. If Column is more confident that Row will cooperate, she is incentivised to cooperate herself if the game is well modelled by a Stag Hunt or by a Prisoner's Dilemma with a cost to defection. As such, the clarity of legal commitments allows correlated interaction among cooperators.

Religious commitments such as oaths may also increase the value of $Pc(R1)$ more than an explicit promise since they are often tied to many other observable behaviours. To elucidate, if someone makes a religious oath but then does not abide by more general religious traditions, we would find

her oath less credible. For example, a person who swears by Allah but does not pray or fast in Ramadan would send a less credible signal of commitment. Living in a religious community does not itself commit an agent to behaving in particular ways, since the agent has not voluntarily signalled. However, as mentioned in Chapter 1, group membership can make one the *kind* of person who can swear credibly. The all-encompassing nature of some religious practices mean that opportunities for observing someone's faith are high, especially since many religious practices are demonstrated through public ritual – for example, going to church, having a bar mitzvah, praying in the mosque, or abstaining from worldly pleasures. Though these acts do not directly increase the probability of detecting defection for any particular oath, they do mean that a receiver's credence in the reliability of an agent's oath is greater in virtue of signalled religiosity. Of course, one's action may not be as visible as it will be in a small-scale society, but this is not to say that institutionalised religion does not make one's actions more observable than they would be otherwise.

Religious commitments can also extend the scope of one's potential partners. As mentioned in Chapter 1, group membership does not itself generate commitments, but it provides a common ground in which one's commitments can be communicated and understood. In-groups now not only include one's immediate society but also spatially remote individuals with whom we have no shared experiences apart from their religious background. A credible religious oath can be made between two Catholics regardless of their country of origin and other differences in their cultural group. As such, institutions may, in certain cases, provide a common ground across groups, enabling commitments to be meaningfully employed. Indeed, this is what we see in international bodies such as the European Union which upholds commitments between countries.

However, the most fundamental change to our practice of making and keeping commitments when commitment is institutionalised is that more costly forms of punishment become cheaper and less risky, allowing commitments to be more credible where reputational effects are less effective. Institutionalisation aids not only partner choice but also partner control. Since commitments can be made to be forcibly followed through by means of third-party intervention, the cognitive demands of reputation-gathering are lowered. Thus, again, we see the coevolution of commitment and cooperation. Previous forms of cooperation provided the selective environment for third-party enforcement of commitments to take hold. In this environment, there arises scope to institutionalise commitment, which confers a number of fitness advantages over commitments made between two agents without recourse to third-party policing.

5.3 On deception and punishment

As with our previous forms of commitment, we have evolved a suite of proximate mechanisms to help us identify potential deception and limit our susceptibility to it. Not only this, but as we have seen, some institutionalised commitments have the additional advantage that they are written down or codified more precisely than our previous forms of commitment. Given the clear terms of the commitment, the receiver may be more easily able to identify defection. If she is so able, the sender's perceived probability of defection being detected, $P_R(d)$, also increases. For example, suppose Sandy's lease agreement specifies that she pays rent on the first day of each month. Knowing this, Sandy's perceived probability of her defection being detected should increase as the second day of the month arrives. This is in contrast to commitment in the group hunting of our ancestors. It is likely the requirements of group hunting were not explicitly codified and thus

whether a particular behaviour counted as defection was uncertain. As a result of this codification, the potential for successful deception is lowered.

Of course, detection of deception will depend on the relative complexity of cooperative interactions. There would most likely be a race between the benefits of institutionalising commitment and the costs of increasing complexity in human social worlds. The lease agreement is a relatively simple interaction so detecting defection is easy. In more complex interactions, the explicitness of institutionalised commitments will aid in detecting defection, but it will likely not exceed the ability of detecting defection in small societies where direct observation of one's defection is possible in shared activities. It is possible that institutionalisation nonetheless reduces the impact of the complexity of cooperative interactions through codifying what counts as cooperation or defection, thereby keeping deception within tolerable limits. The pull of deception will also be lowered if the cost of punishment is known and is great. This is precisely what institutionalisation of third-party punishment does for us; the value of c is increased.

Relatedly, third-party policing can provide us with means of preventing defection due to community oversight. In modern society, this is carried out by regulatory bodies. Consider, for example, the unilateral contracts involved in insurance policies. Not only do we have institutions offering promises in this case, but the promise is itself policed by a regulatory institution which ensures that such promises are upheld and penalises the insurance company if they are not. The Financial Conduct Authority in the United Kingdom makes sure consumers are treated fairly in insurance transactions. The Prudential Regulation Authority ensures that insurance companies have enough capital to fulfil their promises by preventative capital risk management. So potential

deception may be limited by prudential oversight. This works by reducing the value of *x,* the temptation payoff, since temptation is risky in the face of regulatory backlash, increasing the means of detection of defection, and hence $P_R(d)$, or increasing the cost *c*.

Institutionalised commitments may also involve a different kind of cost which limits potential deception. While, earlier, commitments were made credible by socially imposed costs, credibility in the case of some institutionalised commitments – notably modern legal contracts and religious oaths – can be enhanced by the cost of the signal itself. In order to make sense of this, let us first make a distinction between two kinds of costs involved in signals of commitment. There are upfront costs undertaken by the sender which self-select for honesty. An example of an upfront cost of this sort is seen in the costly signalling discussed in Chapter 1. Here, dishonest agents cannot *afford* to signal, the cost is intrinsic to the signal itself. In other words, these signals involve costs which are "hard to fake" (Frank 1988). There are also contingent costs – costs that are incurred contingent upon defection. The reputational effects we have discussed so far are an example of contingent costs – one's reputation is affected only if one defects. It is also an example of a *social,* contingent cost, since it is imposed by another party.

In the case of legal commitments and religious commitments, we see *upfront, intrinsic* costs in addition to the familiar social, contingent costs of reputational effects. These upfront, intrinsic costs are undertaken at the making of the commitment, regardless of whether one cooperates or defects, and are imposed by the nature of the commitment itself rather than by the other party in the interaction. The intrinsic cost is what serves to keep commitment signals credible, since only

honest agents will signal.[101] In the cases of commitment via shared activity, linguistic commitment and moralised commitment, signals were primarily kept honest by *social,* not intrinsic cost. In contrast, modern legal and religious commitments will often involve some upfront cost, which only honest agents undertake.

To illustrate, consider hiring a lawyer to draft an agreement between two individuals. The legal fees involved in hiring the lawyer serve as a costly signal which renders the commitment more credible. It signals to another agent that one intends to cooperate by the terms of the agreement and that one expects others to do so, too. An agent would be disincentivised to pay an upfront cost if the intention was to defect and it was sufficiently likely she would be caught. This is not to say one would never undergo such cost – the profit of successful defection may be great – only that upfront costs act to *deter* such defection in the same way that social contingent costs do. Of course, upon defection, one *also* pays a contingent cost and this comes in two forms. First, the agent suffers whatever financial penalty was laid out in her legal agreement. Second, she suffers a tarnished reputation and may not be chosen in future interaction.

Religious commitments also involve upfront intrinsic costs, but these are not exactly costs one needs to undergo to *make* the commitment but rather costs that one needs to undergo to make the commitment *credible*. To make one's religious oath credible, one must demonstrate religiosity in public settings, which can be costly. Consider Muslim prayer rituals. Irons (2003) documents the practices of a travelling group of Yomut men who, in order to pray at the specified five times per day, must find a safe place to dismount their horses, find clean water to perform ablutions, and

[101] It is supposed that the cost of undertaking the commitment is less than the potential gain from committing and cooperating, else there would be no reason to commit.

find local people to inquire about the direction of Mecca. The undertaking is costly as it involves multiple difficult steps and takes time from their travelling, but it signals to nearby strangers and to the rest of the group that their future commitment signals – oaths in the name of Allah – are more credible. Performing a fast is similarly costly. In some sense, these acts serve as upfront, intrinsic costs – if one did not pay these costs, one's religious oath would be less credible. Religious oaths may also themselves involve upfront costs if there is a ritual surrounding oath-taking. They still also involve contingent costs in the form of theocratic punishment and in the form of reputational effects. In sum, there are always social costs in the case of reneged commitments but, in some cases of institutionalised commitments, there are also intrinsic costs.

Deceptive commitment is less frequent when commitments involve hard-to-fake, costly signals. Through the introduction of upfront costs and additional contingent costs imposed by third parties, deceptive commitment is disincentivised. Legal commitments often involve expensive legal fees in drafting agreements. Religious traditions are usually complicated and thus hard to fake for an outsider who has not spent time learning the traditions. Learning about a particular religious doctrine is time-consuming and costly, meaning there are upfront costs to a religious commitment. Furthermore, religious rituals provide extensive opportunities for monitoring another's behaviour, at least in regard to their religious duties (Irons 2003). This makes it easier to detect potential defectors as religious commitment is made credible by frequent, often observable, behaviours.

Of course, not all institutionalised commitments will involve such upfront costs. This is not directly a function of third-party policing but it is a function of the cultural background surrounding such institutions and the way in which they function. Nonetheless, deception will be limited with

the proliferation of institutionalised third-party punishment since, not only do we see all the honesty-sustaining mechanisms we encountered in the previous chapters, but now the cost of punishment is a more effective deterrent. Detection is no longer limited by receiver resources and punishment can be highly detrimental to the welfare of the deceptive agent. Not only this, but clear codification of the terms of the agreement can make instances of deception easier to detect and regulatory institutions can serve to disincentivise potential defection.

5.4 Modern prosocial humans

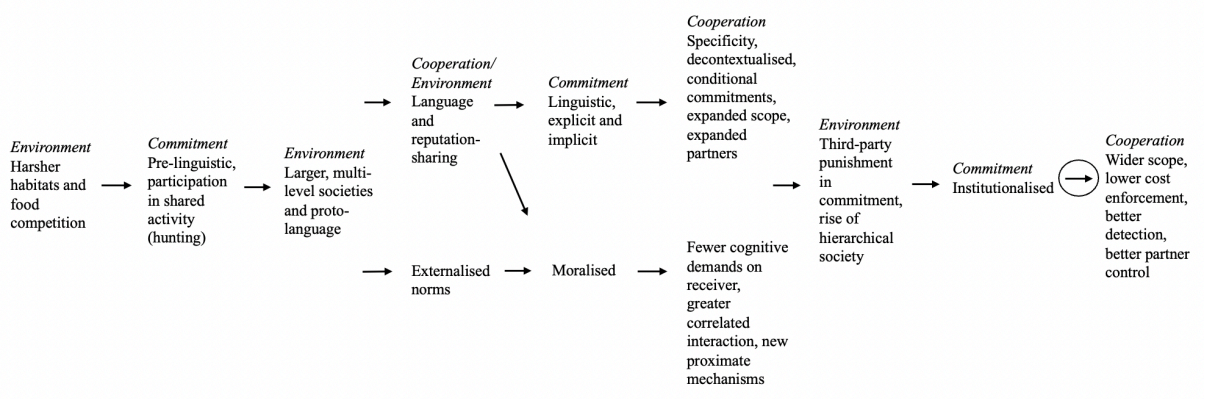

Figure 20: Expanded cooperation as a result of institutionalised commitment.

This chapter has examined the ways in which we have ensured greater clarity, specificity and means of enforcement for our commitments. One particularly effective way of strengthening our commitments has been through institutionalisation. Now we must ask: what further cooperative activity do institutionalised commitments enable us to engage in? Here, we reach the final leg of our coevolutionary story. With institutionalised forms of commitment readily available to almost all agents, we have copious means of making credible and low-cost commitments among strangers

with little risk of exploitation. We see numerous examples of prosocial behaviour on a daily basis. While some animals have cooperative social lives too, humans are often spontaneously cooperative to those with whom they have never before interacted and, as the literature on ultimatum games shows, when there are no reputational benefits to be gained. The explanation, I hold, lies in the proliferation of safe means of cooperation over our evolutionary trajectory due to increased means of commitment. The more we are able to safely commit to engage in cooperative activity, the more selection favours those who have a tendency to be cooperators. Institutionalised forms of commitment widen the scope of one's potential partners at the same time as lowering the costs of enforcement relative to the punishment enacted. Since commitment is often institutionally protected in this way, cooperators can further proliferate in the population. As a result, tendencies toward prosociality will become more common, explaining some of the spontaneous helping behaviour we see in modern day society.

Importantly, I have argued that institutionalised commitment *coevolved* with our previous forms of commitment and cooperation. Language allowed us to broadcast explicit or implicit commitments in a public setting where onlookers have the opportunity and incentive to enforce these commitments in order to bolster their own reputation. The Neolithic revolution was only possible on the basis of growth in the group size and a division of labour, the solving of collective action problems, and the extension of cooperative partners that were enabled by commitment via shared activity, linguistic commitment, and moralised commitment, respectively. The increasingly settled, agricultural life characteristic of the Neolithic created selective pressure for a better means of enforcement of commitment. Increasing organisation and strength of third-party punishment would have been incrementally realised as the demands of cooperation became more complex and

reputation was stressed as an enforcement mechanism. At the same time, institutions were becoming increasingly formalised in response to the complexity of cooperative life, and this had implications for our practices of third-party punishment. In particular, it became organised, lower cost, and common knowledge. I argue that third-party punishment of commitment likely became stable in transegalitarian societies where a leader can corral support among other members of the tribe. Punishment becomes less costly relative to the resources of elites, and more organised. With the modern policing bodies of governments and courts, the threat of third-party punishment is a clear constitutive part of an institutionalised commitment and this had great effects for the extent of human cooperation.

Indeed, without the law to enforce contractual commitments our private property would not be as secure, we would not have modern insurance and banking, or international trade. Legal contracts allow us to make commitments with strangers where reputational information about their past behaviour is irrelevant. This is because if, for example, restitution of property is specified in the contract, we often have little to lose by entering into a commitment and facing potential defection – the court system will ensure that the receiver is paid her due. Thus, laws allow us to undertake risky promises with little opportunity for exploitation, extending the scope of our cooperative interactions.

I have used legal contracts and religious commitments as my primary examples of institutionalised commitment, but which commitments are salient in which societies will depend on differences in culture. In our WEIRD (Western, educated, industrialised, rich and democratic) world,[102] invoking

[102] See Henrich (2020)

divine punishment is perhaps less credible than signing a contract. Goodenough (2003) provides an overview of different types of externally enforced commitments. He notes that, for some communities, to take the Lord's name in vain is a serious wrong, codified in the Ten Commandments. Oaths were also important in Roman courts, Borkowski writes "the procedure [of relying on oaths] may seem irrational but the taking of a formal oath to the gods was a matter of the utmost solemnity for most Romans. Even the worst rogue would hesitate to perjure himself in such circumstances" (Borkowski 1997: 76). Goodenough (2003) notes the use of hostages as a commitment mechanism was prominent in diplomatic relations in the ancient world, and the modern use of security interests and credit instruments mirror this mechanism. They work by holding hostage one's financial security as a means of ensuring good behaviour. Thus, through cultural divergence or development, we can make sense of the wealth of institutionalised commitment mechanisms we see across the world today. Each have ensured greater clarity, specificity and means of enforcement, which allow us to extend our cooperative enterprises further.

Chapter 6: Conclusion

Many explanations of the evolution of cooperation have been offered in the fields of evolutionary biology, anthropology and game theory, including kin selection, group selection, reciprocal altruism, indirect reciprocity, punishment and pre-play signalling, among others. I have argued that a significant and previously overlooked factor in the evolution of human cooperation is the coevolution of commitment and cooperation. That is, I showed how different methods of undertaking commitments in our evolutionary history have enabled more sophisticated forms of cooperation over time which, in turn, create the selective environment for the evolution of increasingly effective commitments. This explanation is primarily one at the level of ultimate causation rather than proximate causation – it refers to the fitness consequences of behaviours rather than the underlying psychological mechanisms which motivate them, though I have said something on the latter as well. Further, it is an explanation of *mutually beneficial cooperative interaction* rather than of *altruistic* cooperation. In the former, both agents benefit from cooperation whereas, in the latter, the altruistic agent undergoes a cost to herself.[103]

In Chapter 1, I introduced my operationalisation of commitment and situated it within the extant literature on commitment and on the evolution of cooperation. A commitment is a means by which an agent can constrain her future choice. In a strategic interaction, the constraint can be signalled to the agent's partner, changing her expectations of the agent's future action. If one agent commits to cooperation, this increases the probability that a receiver will cooperate in turn, since she has reason to believe that she will not be exploited. Not only this, but the signal itself has consequences

[103] This distinction is discussed in more detail in Chapter 1. See West et al. (2007) for an overview.

for the sender's payoffs both at the ultimate and proximate level. At the ultimate level, the most common consequence is that those who renege on their commitments are likely to lose out on future opportunities for beneficial interaction, resulting in a fitness cost to reneging. At the proximate level, senders of commitment signals are more likely to act in line with their signalled intent as a result of social bonds, obligations to fidelity, fear of potential ostracism, and many more motivations. In precise terms, *a commitment is a pre-play signal in a strategic interaction taken at time t, that increases the sender's payoffs for carrying through option X at time t+n, and increases the receiver's probability of the sender carrying through option X.* This definition represents a refined version of Schelling's (1960) theory of commitment.[104]

Commitments are fitness-enhancing in those cases where option *X* is optimal in the long run for the agent, but the agent's motivation to pursue it may be overridden by short-term considerations. To illustrate, an agent who promises loyalty to her spouse in the face of short-term incentives to cheat will secure the beneficial outcome of long-term cooperation with both her spouse and others who take her to be trustworthy (Frank 1988). The commitment to option *X* works by adding either physical, contractual, emotional or reputational costs to choosing against one's commitment, which incentivises option *X* at time *t+n.* In the case of loyalty to one's spouse, option *X* is incentivised by emotional, reputational, and contractual costs to acting against the agent's commitment. At the same time, the spouse who receives the loyal treatment benefits. In this way, commitment can be individually rational while also promoting mutually beneficial cooperation.

[104] A discussion of other theories of commitment – notably, those of Hirshleifer (1984; 2003), Frank (1988), Han (2013) and Back (2007) – and their relative merits is included in Chapter 1.

Commitment is a distinct mechanism for securing cooperation compared to other theories of the evolution of cooperation. It differs from indirect explanations of the evolution of cooperation which appeal to inclusive fitness gains such as kin selection. It also differs from direct explanations of the evolution of cooperation which appeal to long-run benefits to the agent or benefits which outweigh the costs. In Chapter 1, I argued commitment is not synonymous with pre-play signalling and secret handshake explanations of the evolution of cooperation, since commitments are not only signals but signals which involve a cost to reneging. Commitments do not operate only on the basis of reputational consequences, as in indirect reciprocity models, but can involve physical or financial costs. Pre-play signalling can also be used to proactively manipulate one's reputation. The commitment account is not subsumed by the theory of reciprocal altruism, since the latter does not involve pre-play signalling and cannot be used to secure cooperation in one-shot, simultaneous-move games, whilst commitment can. Finally, commitment signals are not always costly, as in the costly signallling hypothesis.

In this dissertation, I elucidated the role of four different types of commitment in our evolutionary history: commitment via shared activity such as in group hunting (Chapter 2); linguistic promising, both explicit and implicit through gossip (Chapter 3); moralised commitment making reference to externalised norms (Chapter 4), and; institutionalised commitment which involves third-party punishment (Chapter 5). The former three involve reputational costs to choosing against one's commitment while institutionalised commitments are both reputationally and contractually enforced. While many of the examples in the dissertation are promises, the account would apply equally well to threats which secure mutual benefit. Some of these promises are intentional and some are unintentional, as seen in Chapter 3. These commitments, I have claimed, coevolve with

cooperation. That is, commitment makes possible new forms of cooperation which make possible new forms of commitment to facilitate cooperation in a more complicated world. In what follows, I will summarise the thesis provided in earlier chapters. The structure of the coevolutionary story I have presented is given below.

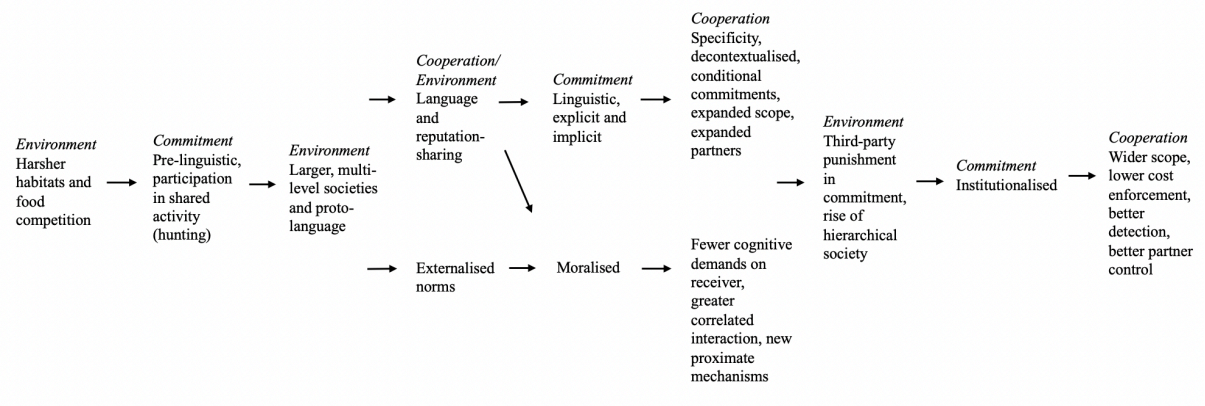

Figure 21: The coevolution of commitment and cooperation.

Our coevolutionary story begins with the ecological changes that made cooperation particularly important for the hominin lineage. Approximately 2 mya, global cooling and drying trends created an expansion of open environments. The change in habitat was coupled with an increase in terrestrial monkeys who outcompeted early hominins for their usual means of subsistence, creating pressure for early hominins to find a new foraging niche. Paleoanthropological evidence suggests that collaborative hunting likely emerged 800 kya with *Homo heidelbergensis* (Tomasello 2012). Many factors were likely important in the evolution of group hunting, notably increasing interdependence and tolerance in scavenging, suppression of the dominance hierarchy due to weaponisation, increasing social intelligence, and increasing impulse control (Tomasello 2012; Sterelny 2012). Note also that cooperative breeding was another early emerging cooperative

activity, selected for as a result of environmental challenges which caused fluctuating food availability, making it difficult to provision young in cooler seasons (Hrdy 2009; O'Connell et al. 1999).

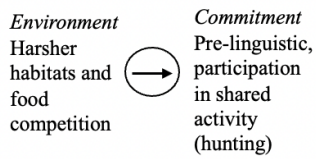

Figure 22: The emergence of pre-linguistic commitment via shared activity.

Group hunting is a cooperative behaviour that is difficult to explain due to the incentive to free-ride off the contributions of others. In Chapter 2, I argued that cooperation becomes directly advantageous to an agent in virtue of a commitment, explaining why defection is limited. As noted in Chapter 1, I use hunting as the illustrative case of commitment via shared activity, though participation in other joint tasks in the ancestral environment may well have possessed the same formal features of commitment. The commitment promotes cooperation because it provides opportunities for beneficial future interaction since the sender can signal one's trustworthiness and follow through. Indeed, Sterelny (2012) suggests that trust is built incrementally through engagement in shared activities and that participation in emotionally-amplified shared experiences, such as hunting, reinforces mutual bonds more so than joint action in calm environments. Reneging on these commitments risks foregoing such opportunities and therefore increases the cost of defection, incentivising cooperation. At the same time, the commitment allows receivers to identify agents as trustworthy or untrustworthy and therefore aids correlated interaction among cooperators in the future.

To expand, the pre-play signal at time *t* is showing up to the hunting excursion – this is a signal which that increases the sender's payoff for pulling his weight in the hunt (option *X*), both at the ultimate level and proximate level. Pulling one's weight at time *t+n* has become the more attractive option since to defect, and lag behind, risks fracturing social bonds and thus not receiving the fruits of the hunt as well as foregoing opportunities for future beneficial interaction, such as in care of offspring. The cost of defection is driven up by the commitment signal, incentivising cooperation. There is evidence in the ethnographic literature of defection from hunting resulting in ostracism (Gurven 2004; Griffin 1982; Balikci 1970). Furthermore, there is evidence of hunting-related activities, notably food-sharing, generating future opportunities for beneficial interaction (Kaplan et al. 2005; Myers 1988; Aspelin 1979). Given that there are future opportunities for beneficial interaction and a cost to defection in real life, group hunting practices align well with the commitment framework. Indeed, even if we do not want to rely on the ethnographic literature, I have argued all that is needed for showing up to the group hunt to count as a commitment is that those who signal cooperation and defect are *less preferentially chosen* in future interactions. Punishment on this account amounts to partner choice and this has a deep history.

Importantly, the evidence suggests those who do not signal that they will hunt (women and children) are not punished by exclusion from the spoils or future interaction should they not participate in the hunt. Furthermore, there is no evidence that men who stay behind at the camp and therefore do not signal there intended participation are punished either. Yet there is evidence that men who go on the hunting excursion and do not pull their weight are punished. This suggests that what drives up the cost of defection is the signal itself, rather than expectations based on

cultural norms or sex-based differences that men in the group ought to participate in the group hunt. This is not to say there were no other factors in maintaining the stability of group hunting – kin selection and cultural transmission will have also played roles – but commitment offers a good model of how stability can be maintained among non-kin and in the short-term, and it is an explanation which fits the empirical literature.

The possibility of free-riding is what incentivises an agent to signal in the first instance. Free-riders hope to gain a greater portion of the captured game without contributing their fair share. In spite of the pull of defection, the honesty of these commitment signals is maintained by *social,* if not intrinsic, costs. Successfully faking participation in hunting activity may not be intrinsically difficult, meaning the probability of detecting defection could be low. However, if the lives and well-being of individuals are intimately connected, as they typically are in hunter-gatherer societies, social exclusion is very costly. Furthermore, social exclusion is widely documented as a form of low-cost punishment. Given this, it is likely the problem of deception did not undermine commitment signalling in the case of shared activity. Though punishment in the form of exclusion from public goods may be more difficult to explain, social exclusion from future, even different, interactions is all that is required to render showing up to the hunt a commitment signal.

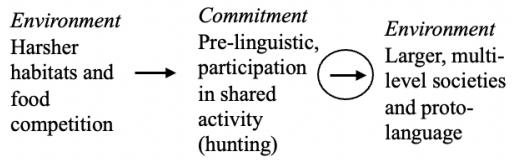

Figure 23: The emergence of larger, multi-level societies.

I have argued the cumulative cultural learning that came with weapon-making and food preparation as a result of cooperative hunting led to the development of clothing, shelter, and more sophisticated tools. Division of labour and specialisation in hunting and gathering skills permit the group to grow due to increased means of subsistence, and to become more productive, in turn favouring the creation of structured environments in which children can learn ever more skills from adults (Sterelny 2012). Due to cooperative breeding, the developmental stages of an infant were informationally richer and more robust to perturbances since there are more sources of information. Not only this, but mothers can devote more energy to producing more babies if their previous young are provisioned. Such developments contributed to extended lifespans and growth of the group size into larger, multi-level networks. It is even proposed that hunting itself was the cause of the group fracturing into a multi-level structure, due to the increase in foraging distance (Layton et al. 2012). So, in Chapter 2, we see how commitment via shared activity contributed to the development of a new selective environment (expansion of the group) in which our next forms of cooperation and commitment would evolve, beginning with the emergence of language.

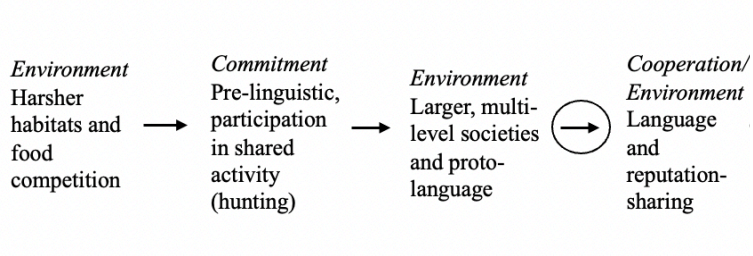

Figure 24: The emergence of language and reputation sharing.

In Chapter 3, I argued that shared activities such as hunting would have played a large role in the emergence of our proto-language and language, drawing on the work of Tomasello (2014).

Specifically, pointing and pantomiming was a useful way of communicating between agents, not only requestively, as seen in great apes, but also informatively, providing the other agent with the information required to complete a joint task. The cognitive capacities that make informative communication possible are what Tomasello (2014) refers to as "joint intentionality" and involve the ability to engage in perspectival representations, to make socially recursive inferences, and to self-monitor. This form of communication was effective when agents shared a common ground pertaining to a joint task, as in hunting, since it is this common ground which enables such gestures to be interpreted.

As we know, these activities contributed to growth of the group size. With more opportunities to interact with agents who do not share a common ground, we see selective pressure for the emergence of a more flexible and decontextualised form of communication than gestural communication – language. Importantly, the cognitive capacities that make abstract, decontextualised communication possible are built upon those developed in our protolanguage as a result of shared activities. We must be able to adopt an agent-neutral perspective or schematise certain situations in an agent-neutral way, using a culturally-constructed representational system, and to govern our thinking and communication by the normative standards of the group. Note that the cognitive capacities that Tomasello cites as crucial to the operation of these two forms of communication (proto-linguistic and abstract and fully decontextualised) would be present regardless of whether we believe the protolanguage was gestural or vocal and regardless of whether these cognitive capacities evolve together or separately. The key claim is that joint tasks select for certain cognitive capacities for communication which lay the groundwork for abstract and decontextualised language.

As such, cooperation in the form of information-sharing via language coevolved with commitment via shared activity. Commitment made possible stable cooperative hunting, which enabled the formation of larger groups, and selected for the use of language in these groups. The second of these coevolutionary claims is *causal* – joint tasks were a driver of language evolution – while the first claim holds that shared activities dependent on commitment provided the cognitive *preconditions* for a new form of cooperation. Language allows for cooperative activities to be undertaken with strangers with reduced risk of exploitation, given that information can now be shared and accessed about a potential partner's reputation.

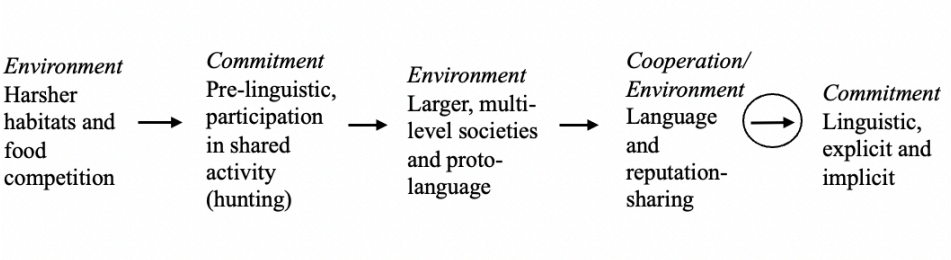

Figure 25: The emergence of linguistic commitment.

This cooperative environment set the stage for a new form of *commitment* to evolve – linguistic promising. Of course, there are also intermediate stages between pre-linguistic commitment and commitment dependent on fully abstract and decontextualised language. Indeed, commitments will become more sophisticated and effective as our protolanguage develops in flexibility and precision. That is, our language capacities likely coevolved with the explicitness and efficacy of commitment.

Linguistic commitment operates in the same basic way as commitment via shared activity. That is, it alters sender payoffs via increasing the cost of defection, since reneging risks potential

exclusion from current and future benefits. It also allows the receiver to identify trustworthy partners, aiding correlated interaction. I suggested explicit promising may have arisen out of the pressure to make explicit the roles of interaction in shared activity. In tasks involving a division of labour, what is needed in complicated joint activities is a coordination device or symmetry breaker (Skyrms 1996). As we engaged in the more sophisticated tool-making and projectile weaponry of the Middle Stone Age, I suggested there arose selective pressure to communicate one's role in a joint activity, where natural salience alone may not suffice. Once these roles can be articulated, they form expectations which will result in negative consequences if they are not met. This combination of signalling and altering of one's payoffs is characteristic of commitment.

With explicit linguistic commitments, the pre-play signal at time *t* is the utterance of the promise, and the option *X* for which one's payoff has changed is the cooperative matter on which one has promised. The utterance increases the sender's cost of defection since reneging on a commitment makes one vulnerable to exclusion from interaction, thereby changing one's fitness consequences. Opportunities for future interaction now exist in virtue of the interrelated practices of the group and the fact that we have access to the reputation of others, making them potential partners for cooperative interaction. At the proximate level, explicit promising is likely incentivised by fear of ostracism for which there is ample psychological evidence (Williams 2007; Emler 1990).

Another form of linguistic commitment is implicit promising via gossip. In an environment where reputation sharing is enabled by language, we can commit to norm adherence by passing judgment on others who have or have not so adhered. Utterances of the form "it is terrible that Danny did not attend the fortnightly ritual dance" become commitments not to behave in like manner. Such

judgments are taken to be commitments as long as the receiver of the signal believes the sender to be part of the same cultural group as the agent about whom they are gossiping, and to occupy the relevant social role. For example, the utterance would count as a commitment if the utterer is of the social status that makes her relevantly obligated to attend the fortnightly ritual dance. Gossip is a means of reputation sharing while also being a commitment, making it rich with information for securing cooperation.

I suggested the origin of implicit commitment is to be found in the value of information-carrying signals for correlating interaction. Agents who were able to infer another's future course of action on the basis of her gossip – even by initial accident – may have received more cooperative benefits than agents who did not. They not only shared and received reputational information about others but simultaneously used this information about attitudes toward norm adherence or violation to advertise trustworthiness to others and to discriminate among potential partners. An initial correlation device can become a commitment as long as there is a cost to reneging. Through this form of commitment, we can extend the scope of cooperation further. Linguistic commitment affords us even more opportunities for assessment of reliable partners, aiding correlated interaction.

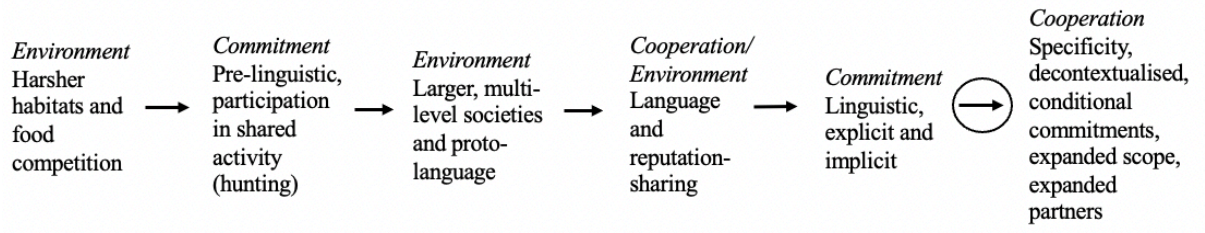

Figure 26: Expanded cooperation as a result of linguistic commitment.

Linguistic commitments, both explicit and implicit, offer a number of fitness advantages over commitment via shared activity. Language is decontextualised, allowing us to make commitments in multiple contexts where a shared common ground does not exist. It also has scope to be more specific than non-linguistic signals of cooperation, forming more precise expectations of future behaviour. Linguistic commitments allow us to make conditional commitments and therefore extends the scope of commitment to Trust games. We may also specify the penalties of reneging, which can make commitments which were not previously credible become credible, since reputational damage may not only come from reneging on the commitment itself but also reneging on the specified penalty, compounding the costs of defection. With linguistic promising, we can make commitments to reciprocal exchange over time. Finally, language allows for apology and amendment more effectively than non-linguistic communication, allowing us to recover our reputation in the face of a reneged commitment.

Importantly, commitments no longer depend on opportunities for shared activities or on social bonding. Reputation sharing provides means for commitment to be made among unfamiliar partners, since previous interaction is not required to determine a potential partner's trustworthiness if one has access to their reputation. Information from multiple sources will aid reliability here. Connected networks may even facilitate commitments among in-group strangers – agents about whom one has no information – if there is a possibility that one's reputation may reach others with whom one will interact in the future.

However, linguistic signals are also cheap to fake. What secures honesty, again, is social cost. I argued we have developed proximate mechanisms both to attend to potential deception and to punish it when it occurs. There is evidence that social exclusion is employed as a low-cost form of punishment and that honesty can be maintained when there is a possibility of punishment (Eisenberger et al. 2003; Williams 2007; Archer & Coyne 2005; Molho et al. 2020; Behnk et al. 2018). Credibility is also enhanced by reputation sharing in the community. News of one's defection will spread to other potential partners, increasing the cost of defection. It is also worth noting that the pull of deception will be greater when the cooperative enterprises we engage in are few and far between, and when the stakes of the interaction are high. This is because the benefits of deception will be greater and there is little to lose if the agent does not expect to interact with the same network of individuals again. However, in larger groups where information-sharing is rife and cooperative activities are more varied than we have seen in hunter-gatherer societies, it is likely that opportunities for beneficial interaction arise frequently and include many lower-stakes interactions, where the pull of deception is low.

New forms of cooperation coevolved with this expanded means of commitment. After the advent of language, it was possible to make commitments precisely among new partners, commitments that were sophisticated and flexible enough to apply to new contexts, and commitments to reciprocation over time. Given the wealth of opportunities for employing linguistic promising, agents are afforded more opportunities to observe another's behaviour and thus infer their reliability, extending the scope of safe cooperation yet further. In-group strangers become potential cooperative partners, as long as there is the possibility that news of one's actions reach others in one's social network, even one's *potential* social network. Furthermore, new forms of cooperation

that were not possible before become so. For example, pooling funds for a building investment. Here, the ability to abstractly refer and precision about the division of labour in financial and physical investment would aid in achieving mutual benefit, but what is required to overcome the incentive to free-ride is the ability to make credible commitments to investment and for those commitments to be believed.

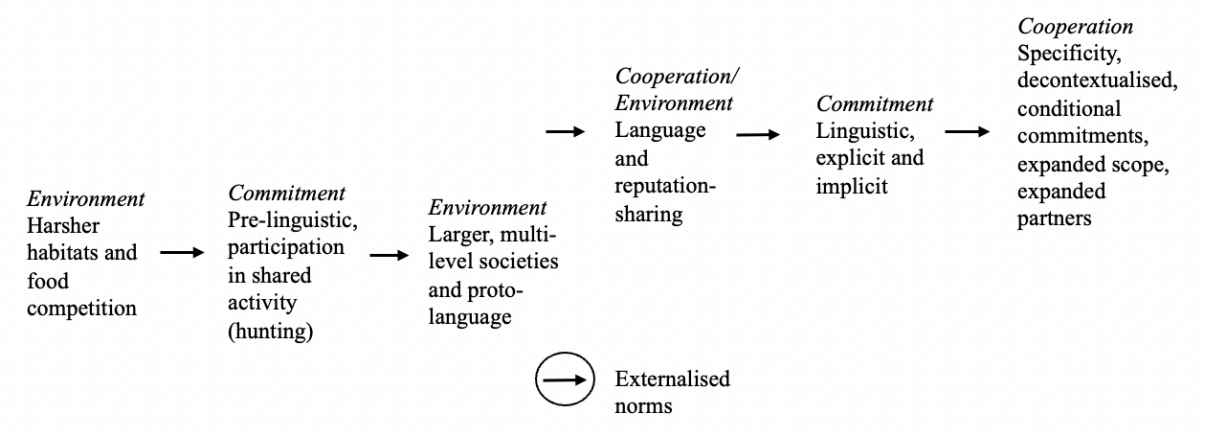

Figure 27: The emergence of externalised norms.

In Chapter 4, I argued shared activity which depended commitment also contributed to the cooperative environment necessary for a third form of commitment to evolve: moralised commitment. I use the term "moralised commitment" to refer to those commitments based on externalised norms. To say that a norm is externalised is to say that it is experienced as imposed on us from the outside and exacting a demand on all, regardless of their group (Stanford 2018). Though not all moral norms are externalised, many typically moral norms are, so I have used evidence concerning moral norms as a proxy for externalised norms. Though we cannot determine exactly whether non-moralised linguistic commitment preceded moralised linguistic commitment, we do know that commitment via shared activity was a precursor to both. In particular, hunting

and other shared activities, such as cooperative breeding, coevolved with the selective environment that made moralised commitment possible – these forms of commitment contributed to both the precursors and selection pressure for norm externalisation.

I suggested that our moral cognition has its foundations in the perspective-taking capacities and prosocial emotions that coevolved with commitment-based shared activity (Goldman 1993; Gordon 1995; Nichols 2004). That is, the empathy typically characteristic of moral cognition requires taking on another's perspective, and such capacities likely evolved to facilitate action in joint tasks such as hunting and alloparenting. The affective mechanism which motivates helping behaviour is also strengthened by the effect alloparenting and hunting has on perceptually tuning us to emotions (Hrdy 2009). Furthermore, externalisation of norms became an important mechanism for partner choice in the context of increasing group size, which was made possible by early hominin shared activities. This is particularly so as opportunities for shared activity and access to reputation decrease as the size of the group grows, meaning our previous partner choice mechanisms become less effective. Externalised norms correlate interaction by establishing a connection between an agent's own motivation to adhere to a certain norm and her means of choosing reliable partners, those who share her views (Stanford 2018). As such, there is again a coevolutionary relationship between forms of commitment and the cooperative environment that they select for, which makes new forms of commitment possible. The first of these coevolutionary claims is, again, about the cognitive *precursors* that needed to be in place for a new kind of cooperation to evolve – correlated interaction on the basis of externalised norms has its precursors in perspective-taking and prosocial emotions, since these are the foundations of moral cognition.

The second claim is causal, that the larger group enabled by commitment via shared activity *selected for* externalised norms as a means of correlating interaction.

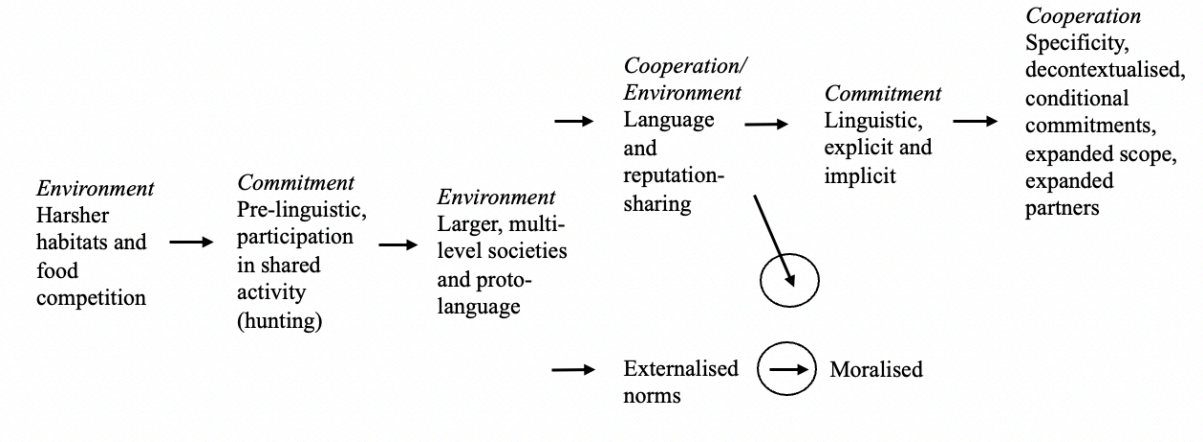

Figure 28: The emergence of moralised commitment.

Moralised commitments can be direct, as in assertions of the form "doing *X* is wrong" or indirect, in gossip about the adherence or violation of an externalised norm by another agent, in fables and in story-telling. All operate in the same basic way. If the receiver of the signal believes the sender takes the norm to be externalised, she holds the sender to the standard she has espoused, since externalised norms apply to all individuals irrespective of group. Not only this, but she may be prepared to exclude this individual from further interaction based on her lack of adherence to her judgment. Our unique and powerful attitude toward hypocrisy is what sustains this form of commitment and imposes a cost to reneging on the sender (Powell & Smith 2013; Alicke 2000; Shklar 1984; Jordan et al. 2017). Moralised commitments therefore work in like manner to those commitments we have seen previously – by changing sender payoffs and receiver expectations. They are also kept honest via the same socially imposed costs and proximate attitudes toward promise-breaking that serve to keep non-moralised linguistic commitment evolutionarily stable.

Further, if the sender in fact takes the norm to be externalised, she is internally motivated (at least on many accounts of moral motivation) to act in line with her signalled intent, independently of any potential ostracism she may face for not doing so. I have argued that potential deception is limited by our attitudes toward hypocrisy as well as our attitudes toward externalised norms. Not only this, but the frequency with which we exchange moral gossip will aid in assessment of reliable partners, since commitments are being made in the ordinary course of reputation sharing (of the right sort).

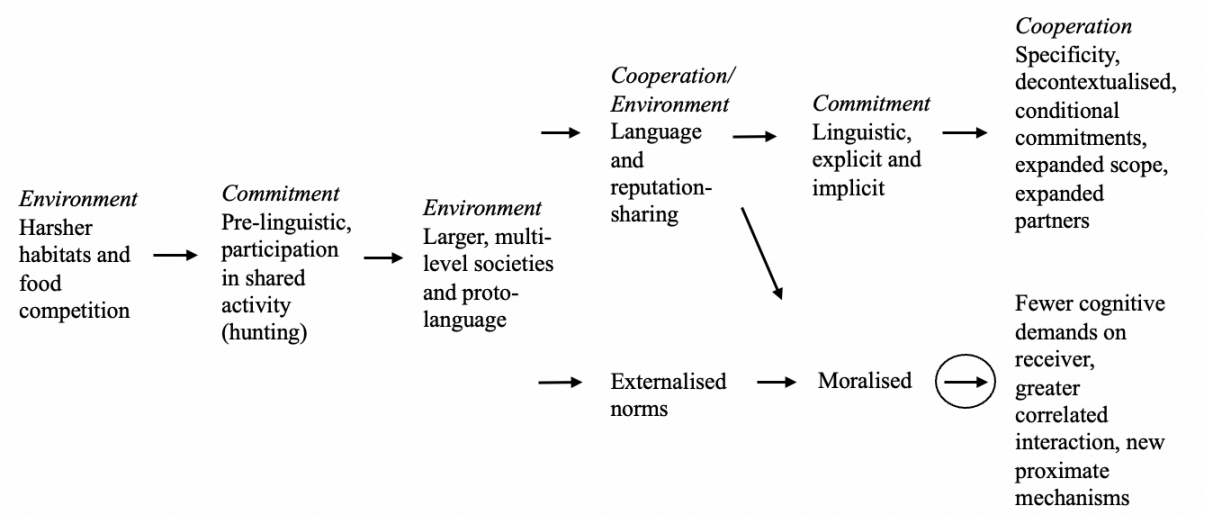

Figure 29: Expanded. cooperation as a result of moralised commitment.

Moralised commitments have fitness advantages over non-moralised commitments. In particular, gossip about the behaviour of another party toward an externalised norm constitutes a commitment irrespective of whether the sender shares the same cultural background and social role as the agent about whom they gossiped. They therefore require fewer inferences on the part of the receiver, meaning commitments are easier to make and easier to identify. Not only this, but by either directly expressing one's attitude about an externalised norm or by gossiping about the behaviour of a third party with respect to an externalised norm, the sender reveals information about how they would

treat *others* who behaved this way, including the receiver. By making one's expectations of one's interactive partner clear, the sender of the moralised commitment better secures correlated interaction with like types. Finally, psychological evidence points to the efficacy of moralised language in changing receiver expectations of the sender's future course of action, thereby making commitments more credible. Part of the explanation is the unique phenomenology of externalised norms which involves new and powerful proximate mechanisms related to guilt and shame. So, we add the following to our coevolutionary story: commitment via shared activity and language set the stage for the development of externalised norms which, in turn, provided the resources to extend the scope and power of our commitments.

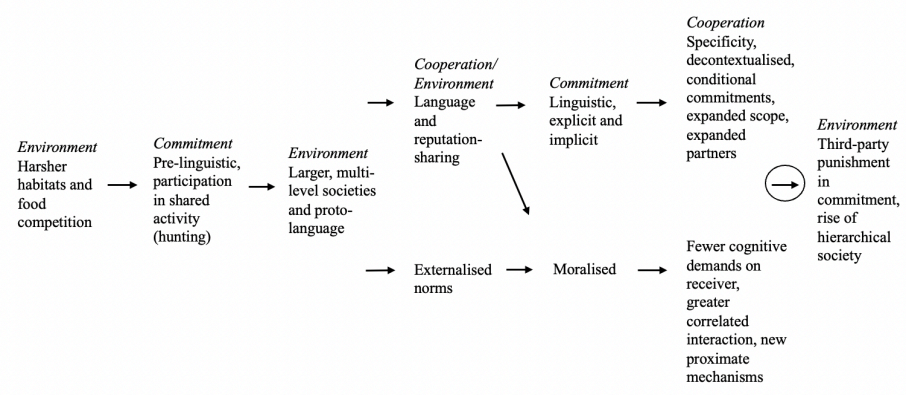

Figure 30: Third-party punishment of commitments and the rise of hierarchical societies.

In Chapter 5, I suggested that the emergence of third-party punishment of commitments was a result of linguistic commitments made in public settings. In these settings, onlookers have the opportunity to enforce the commitment and may be motivated to do so to signal their trustworthiness to their own subgroup. Our previous forms of commitment and cooperation also laid the groundwork for the Neolithic revolution in which this third-party punishment would be strengthened. In particular, hunting and other shared activities allowed innovation and a division

of labour which we know contributed to a growth of the group size and the increasing complexity of our cooperative interactions. Linguistic commitment allowed us to coordinate our actions in more complex collective action problems which demand plasticity of response. Moralised commitment extended cooperation more safely to strangers and introduced a new proximate mechanism for ensuring agents follows through on their commitments – the phenomenology of externalised norms. Yet, as we transitioned to a more settled, agricultural life in the Neolithic, we see that interaction with strangers was becoming more common just as tasks were becoming more complex. This created selective pressure for a more effective means of enforcement for commitment, over and above reputational effects. One such mechanism is third-party punishment. With the rise of hierarchical societies shortly after the Neolithic revolution, we had more resources for the deployment of third-party punishment. In particular, elites could more easily garner material support. At the same time, institutions were becoming more formalised – clearly codified, impersonal and organisational. This had further consequences for our practices of third-party punishment. With the rise of institutions, third-party punishment became organised, more precise, and lower cost. All of this culminates in the development of institutionalised commitment.

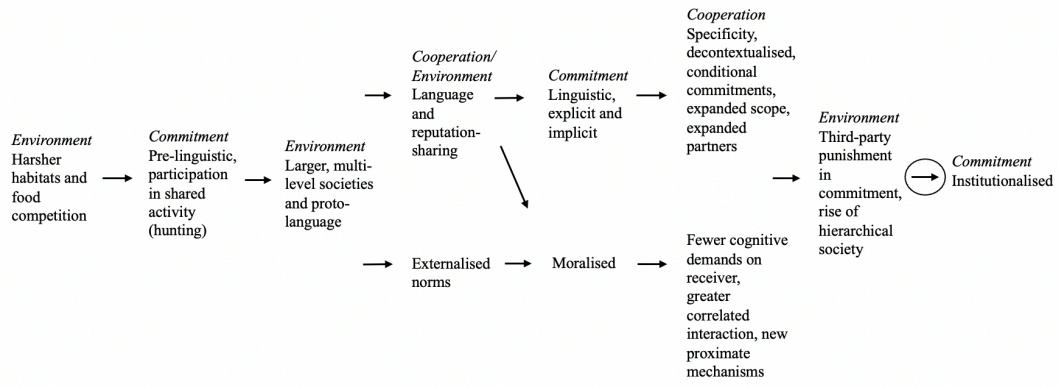

Figure 31: The emergence of institutionalised commitment.

Institutionalised commitments are those commitments enforced by third-party punishment conducted in an organised manner, where it is common knowledge that transgressive behaviours will be punished and, typically, where the punishment is carried out by a body which undertakes a policing role in the community and is funded by its members. Modern examples are rife, involving any sort of contractual agreement backed by law. Yet institutionalised commitment also comes in degrees. The case of the Turkana community punishing those who renege on their commitments to war raids represents a case where punishment is enacted by third parties and is guided by norms of enforcement, but there is no one policing body (Mathew & Boyd 2011). Institutionalised commitments again change sender payoffs at both the ultimate and proximate level and change receiver expectations of the sender's action. However, they differ from our previous forms of commitment in that, now, the incentive to adhere to one's commitment not only comes from reputational enforcement, but also contractual enforcement. If an agent enters into a legal contract, she becomes more vulnerable to penalties imposed by others apart from the person to whom she fails the commitment. Third parties are also responsible for imposing a cost for defection. In addition to reputational consequences, a third-party such as the state may impose a fine, imprisonment or reparations equivalent to fulfilment of the commitment. Similar mechanisms are at play with religious oaths, too.

Along with the advent of institutionalised commitment, there came a new means of protecting oneself from deception. Unlike the commitments of our previous chapters, some legal and religious commitments involve upfront, intrinsic costs. That is, costs which are undertaken when the commitment is issued, even if one does not defect, and in virtue of the commitment itself, rather

than imposed by another party. The legal fees involved in writing a contract and the religious demonstrations needed to make religious commitments credible fall on both honest and deceptive agents alike. This upfront cost deters defectors from signalling if there is a decent chance of being caught reneging. As a result, in such cases, it is likely that only honest agents are the ones who can afford to signal. Furthermore, many institutionalised commitments also benefit from institutionalised safeguards to prevent defection. Legal contracts and insurance policies are managed by regulatory bodies to ensure that commitments are carried through. This lessens the pull of deception, making commitment signals more likely to be honest. Of course, the previous honesty-sustaining mechanisms also apply, in particular, socially imposed costs in the form of a tarnished reputation.

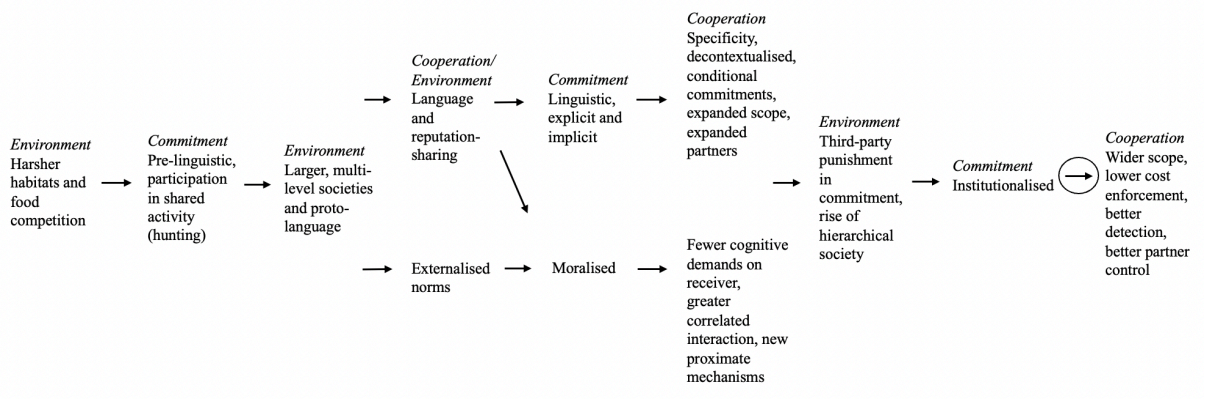

Figure 32: Expanded cooperation as a result of institutionalised commitment.

Institutionalised commitments offer further fitness advantages over previous forms of commitment. They can be made clearer, more effectively altering receiver expectations of the sender's action. It is often easier to identify defection given the clarity of the terms. Most importantly, third-party enforcement can impose costly punishment on behalf of the wronged

agent, increasing the cost of defection relative to the cost of punishment. This can even take the form of ensuring that the commitment is followed through, for example, in sales. This means we have not only a partner choice mechanism in commitment but also a partner control mechanism – something that ensures that defectors cooperate in future rounds. These mechanisms can operate in the absence of reputational consequences and so provide us with means of extending our trust to agents who do not share our social network, lowering the cognitive demands of reputation-gathering since commitments are often institutionally protected. In addition, with religious commitments (and some legal commitments) the scope of one's in-group is expanded beyond geographical boundaries, the credibility of religious oaths can be backed by many frequently observable and public religious behaviours, and the motivation to follow through involves new and powerful proximate mechanisms.

The effect that these new forms of commitment had on cooperation is clear. Without the law to enforce contracts, we would not have had the proliferation of private property we see today, modern insurance and banking, or international trade. Furthermore, if legal contracts allow us to make and believe commitments among partners about whom we do not possess reputational information, this lowers the cognitive demands of trust. Restitution of property can be specified in the contract, meaning there is nothing to lose from potential defection. When prosociality can be institutionally protected in this way, we have an explanation of the proliferation of prosocial tendencies we see today, in the spontaneous helping behaviours of complete strangers. The suggestions is that these cooperative tendencies have spread in the human population because cooperation was safeguarded by credible commitments and a means of enforcing these commitments.

When exactly each of these forms of commitment evolved is less important than what they enabled us to do. They serve to provide an effective set of tools for securing mutually beneficial outcomes when defection may be individually optimal and have emerged in response to the increasing complexity of cooperative interactions. Commitments secure cooperation by altering the fitness costs and benefits of defection relative to maintaining opportunities for interaction in the face of potential backlash. They can therefore secure the long-run optimal option for the sender of the commitment signal and act as a correlation mechanism for the receiver of the commitment signal. Over our evolutionary history, each form of commitment has changed our cooperative environment in ways that then select for the evolution of increasingly effective commitments in response to new challenges. Not only this, but the commitment account is a good explanation of many of our actual cooperation-sustaining mechanisms. I have presented a large body of anthropological and psychological evidence that such commitment practices were widespread in our evolutionary history and that they continue to be present, and motivationally powerful, today.

REFERENCES

Alexander, R. D. 1987. A Biological Interpretation of Moral Systems. *Zygon 20*(1), 3-20.

Ambady, N., & Rosenthal, R. 1992. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin 111*(2), 256–74. *America 98*(13), 6993 – 6996.

Archer, J., & Coyne, S. M. 2005. An integrated review of indirect, relational, and social aggression. *Personality and Social Psychology Review 9*(3), 212-30.

Arnold, K., & Aureli, F. 2007. Postconflict reconciliation. In C. Campbell et al. (eds), *Primates in perspective*, 592–608. Oxford: Oxford University Press.

Aspelin, P. 1979. Food distribution and social bonding among the Marmaindê of Mato Grosso, Brazil. *Journal of Anthropological Research 35*(3), 309–327.

Axelrod, R. 1984. *The Evolution of Cooperation.* New York: Basic Books.

Axelrod, R., & Hamilton, W. D. 1981. The Evolution of Cooperation. *Science, New Series 211*(4489), 1390-1396.

Back, I. 2007. *Commitment and Evolution: Connecting Emotion and Reason in Long-term Relationships*. ICS Dissertation Series 133.

Balafoutas, L., Nikiforakis, N., & Rockenbach, B. 2014. Direct and indirect punishment among strangers in the field. *Proceedings of the National Academy of Sciences 111*, 15924–15927.

Balikci, A. 1970. *The Netsilik Eskimo.* Natural History Press.

Barclay, P., & Willer, R. 2007. Partner choice creates competitive altruism in humans. *Proceedings of the Royal Society B: Biological Sciences 274*(1610), 749-753.

Barrett, J. A., & Skyrms, B. 2017. Self-assembling Games. *British Journal for the Philosophy of Science 68*(2), 329-353.

Bauers, K. A. 1993. A functional analysis of staccato grunt vocalizations in the stumptailed macaque (Macaca arctoides). *Ethology 94*(2), 147–161.

Behnk, S; Barreda-Tarrazona, I., & García-Gallego, A. 2018. Punishing liars—How monitoring affects honesty and trust. *PLoS ONE 13*(10), 1-30.

Berg, J.; Dickhaut, J., & McCabe, K. 1995. Trust, Reciprocity, and Social History. *Games and Economic Behavior 10*(1), 122-142.

Bergmann, J. R. 1993. *Discreet indiscretions: The social organization of gossip.* New York: Aldine de Gruyter.

Bicchieri, C. 2006. *The Grammar of Society: The Nature and Dynamics of Social Norms.* New York: Cambridge University Press.

Bickerton, D. 1990. *Language and Species*. Chicago: University of Chicago Press.

Bingham, P. 1999. Human Uniqueness: A General Theory. *Quarterly Review of Biology 74*(2), 133– 169.

Bingham, P. 2000. Human Evolution and Human History: A Complete Theory. *Evolutionary Anthropology 9*(6), 248– 257.

Blair, R. J. R. 1999. Psychophysiological responsiveness to the distress of others in children with autism. *Personality and Individual Differences 26*(3)*,* 477–85.

Bliege Bird, R.; Smith, E., & Bird, D. 2001. The Hunting Handicap: Costly Signaling in Human Foraging Strategies. *Behavioral Ecology and Sociobiology 50*(1), 9-19.

Boles, T.; Croson, R., & Murninghan, K. 2000. Deception and retribution in repeated ultimatum bargaining. *Organ, Behavior and Human Decision Processes 83*, 235–259.

Borkowski, A. 1997. *Textbook on Roman Law.* London: Blackstone Press.

Bowles, S., & Gintis, H. 2004. The evolution of strong reciprocity: cooperation in heterogeneous populations. *Theoretical Population Biology 65*(1), 17-28.

Boyd, R.; Gintis, H.; Bowles, S., & Richerson, P. 2005. The Evolution of Altruistic Punishment. In H. Gintis, S. Bowles, R. Boyd & E. Fehr (Eds.) *Moral Sentiments and Material Interests: The Foundations of Cooperation in Economic Life,* pp. 215-227. Cambridge, MA: MIT Press.

Brandts, J., & Charness, G. 2003. Truth or Consequences: An Experiment. *Management Science 49*(1), 116–130.

Brown, W. M. 2003. Are there nonverbal cues to commitment? An exploratory study using the zero-acquaintance video presentation paradigm. *Evolutionary Psychology 1,* 42–69.

Brusse, C. 2020. Signaling theories of religion: models and explanation. *Religion, Brain & Behavior 10*(3), 272-291

Bunn, H. T., & Pickering, T. R. 2010. Bovid mortality profiles in paleoecological context falsify hypotheses of endurance running-hunting and passive scavenging by early Pleistocene hominins. *Quaternary Research 74*, 395-404.

Burling, R. 2011. Words came first: adaptations for word-learning. In M. Tallerman & K. R. Gibson (Eds.), *The Oxford Handbook of Language Evolution,* pp. 406-16. Oxford: Oxford University Press.

Byrnes, D. A., & Kiger, G. 1988. Contemporary measures of attitudes toward Blacks. *Educational and Psychological Measurement 48*(1), 107–118.

Chapais, B. 2008. *Primeval kinship: how pair-bonding gave birth to human society.* Cambridge, MA: Harvard University Press.

Charness, G., & Dufwenberg, M. 2006. Promises and Partnership. *Econometrica 74*(6), 1579–1601.

Chen, Z.; Williams K. D.; Fitness, J., & Newton, N. C. 2008. When hurt will not heal: exploring the capacity to relive social and physical pain. *Psychological Science 19*(8), 789-95.

Chomsky, N. 1980. *Rules and Representations*. London: Basil Blackwell.

Clark, H. 1996. *Uses of Language.* Cambridge: Cambridge University Press.

Clark, R., & Kimbrough, S. O. 2017. Social structure, opportunistic punishment and the evolution of honest signaling. *PLoS ONE 12*(12).

Colwell, R. K. 1981. Group selection is implicated in the evolution of female-biased sex ratios. *Nature 290*(5805), 401–404.

Crandall, C. S. 1991. Multiple stigma and AIDS: Medical stigma and attitudes toward homosexuals and IV-drug users in AIDS-related stigmatization. *Journal of Community and Applied Psychology 1*(2), 165–172.

Currie, T. E.; Turchin, P.; Bednar, J.; Richerson, P. J.; Schwesinger, G.; Steinmo, S.; Wacziarg, R., & Wallis, J. 2016. Evolution of Institutions and Organizations. In D. S. Wilson & A. Kirman (Eds.), *Complexity and Evolution Toward a New Synthesis for Economics,* pp. 199-234. Cambridge, MA: MIT Press.

Damon, W. 1977. *The social world of the child.* San Francisco: Jossey-Bass.

Darwall, S. 2006. *The second-person standpoint: Morality, respect, and accountability*. Cambridge, MA: Harvard University Press.

Darwin C. 1871. *The descent of man, and selection in relation to sex.* London, UK: Murray.

Dawkins, R. 1976. *The Selfish Gene*. New York: Oxford University Press.

Dawkins R., & Krebs J. R. 1978a. Animal Signals: Information or Manipulation? In J. R. Krebs & N. B. Davies (Eds.), *Behavioural Ecology: An Evolutionary Approach,* pp. 283-309. Oxford: Blackwell.

Dawkins R., & Krebs J. R. 1978b. Animal Signals: Mind-Reading and Manipulation. In J. R. Krebs & N. B. Davies (Eds.), *Behavioural Ecology: An Evolutionary Approach,* pp. 381-402. Oxford: Blackwell.

De Matta, A.; Gonçalves, F. L., & Bizarro, L. 2012. Delay discounting: concepts and measures. *Psychology & Neuroscience 5*(2).

De Waal, F. B. M. 1996. *Good Natured: The Origins of Right and Wrong in Humans and the Other Animals*. Cambridge, MA: Harvard University Press.

De Waal, F. B. M., & Van Roosmalen, A. 1979. Reconciliation and Consolation among Chimpanzees. *Behavioral Ecology and Sociobiology 5*(1), 55-66.

Dediu, D., & Levinson, S. C. 2013. On the antiquity of language: the reinterpretation of Neandertal linguistic capacities and its consequences. *Frontiers in Psychology 4:* 1-17.

Deigh, J. 1995. Empathy and universalizability. *Ethics 105*(4), 743–63.

Ditrich, L., & Sassenberg, K. 2016. It's either you or me! Impact of deviations on social exclusion and leaving. *Group Processes & Intergroup Relations 19*(5), 630-652.

Dunbar, R. I. M. 1993. Coevolution of neocortical size, group size and language in humans. *Behavioral and Brain Sciences, 16*(4), 681-694.

Dunbar, R. I. M. 1996. *Grooming, Gossip and the Evolution of Language.* Cambridge, MA: Harvard University Press.

Dunbar, R. I. M. 1998. The Social Brain Hypothesis. *Evolutionary Anthropology: Issues, News, and Reviews 6*, 178-190.

Dunbar, R. I. M. 2004. Gossip in evolutionary perspective. *Journal of General Psychology 8,* 100–110.

Dunbar, R. I. M. 2011. The Social Origins of Language. In M. Tallerman & K. R. Gibson (eds), *The Oxford Handbook of Language Evolution*. Oxford: Oxford University Press.

Eisenberg, N., & Fabes, R. 1990. Empathy: Conceptualization, measurement, and relation to prosocial behavior. *Motivation and Emotion 14*(2), 131–49.

Eisenberger, N. I.; Lieberman, M. D., & Williams, K. D. 2003. Does rejection hurt? An fMRI study of social exclusion. *Science 302*(5643), 290–292.

Elster, J. 2000. *Ulysses Unbound: Studies in Rationality, Precommitment, and Constraints.* Cambridge: Cambridge University Press.

Emler, N. 1990. A Social Psychology of Reputation. *European Review of Social Psychology 1*(1), 171-193.

Endicott, K. 1988. Property, Power and Conflict among the Batek of Malaysia. In T. Ingold; D. Riches & J. Woodburn (Eds.), *Hunters and Gatherers 2: Property, Power and Ideology*, pp. 110-127. Oxford: Berg.

Fehr, E., & Fischbacher, U. 2004. Third-party punishment and social norms. *Evolution and Human Behavior, 25*(2), 63–87.

Fernandes, S.; Kapoor H., & Karandikar, S. 2017. Do We Gossip for Moral Reasons? The Intersection of Moral Foundations and Gossip. *Basic and Applied Social Psychology 39*(4), 218-230.

Foley, R., & Gamble, C. 2009. The ecology of social transitions in human evolution. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences 364*(1533), 3267 – 3279.

Foley, R. A., & Lee, P. C. 1991. Ecology and energetics of encephalization in hominid evolution. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences 334*(1270), 223–232.

Forber, P., & Smead, R. 2015. Evolution and the classification of social behavior. *Biology and Philosophy 30*(3), 405-421.

Foster, E. K. 2004. Research on gossip: Taxonomy, methods, and future directions. *Review of General Psychology 8*(2), 78–99.

Frank, R. H. 1988. *Passions Within Reason: The Strategic Role of the Emotions.* New York: Norton.

Frank, R. H. 2003. Cooperation through Emotional Commitment. In R. Nesse (Ed.), *Evolution and the Capacity for Commitment*, pp. 57-76. New York: Russell Sage Foundation.

Frank, R. H. 2011. The Strategic Role of the Emotions. *Emotion Review 3*(3), 252–254.

Freeberg, T. M. 2006. Social Complexity can Drive Vocal Complexity. *Psychological Science 17*, 557–561.

Freeberg, T. M., & Lucas, J. R. 2011. Information theoretical approaches to chick-a-dee calls of Carolina chickadees (Poecile carolinensis). *Journal of Comparative Psychology 126*(1), 68-81.

Freeberg, T. M.; Dunbar, R. I. M., & Ord, T. J. 2012. Social complexity as a proximate and ultimate factor in communicative complexity. *Philosophical Transactions of the Royal Society B: Biological Sciences 367*(1597), 1785–1801.

Fried, C. 1981. *Contract as Promise: A Theory of Contractual Obligation.* Cambridge, MA: Harvard University Press.

Gibbard, A. 1990. *Wise Choices, Apt Feelings.* Cambridge, MA.: Harvard University Press.

Gilbert, M. 2013. *Joint Commitment: How We Make the Social World.* Oxford: Oxford University Press.

Gintis, H. 2011. Gene-culture coevolution and the nature of human sociality. *Philosophical Transactions of the Royal Society B: Biological Sciences 366*, 878–888.

Godfrey-Smith, P. 2014. Sender-Receiver Systems within and between Organisms. *Philosophy of Science 81*(5), 866-878.

Goldin-Meadow, S. 2003. *The Resilience of Language. Essays in Developmental Psychology*. New York: Psychology Press.

Goldman, A. I. 1993. Ethics and cognitive science. *Ethics 103*(2), 337–60.

Goodenough, W. H. 2003. Law and the Biology of Commitment. In R. Nesse (Ed.), *Evolution and the Capacity for Commitment*, pp. 262-291. New York: Russell Sage Foundation.

Goodwin, G. P., & Darley, J. M. 2008. The Psychology of Meta-Ethics: Exploring Objectivism. *Cognition 106*, 1339–1366.

Gordon, R. 1995. Sympathy, simulation, and the impartial spectator. *Ethics 105*(4), 727–42.

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

Grégoire, P., & Robson A. 2003. Imitation, group selection and cooperation. *International Game Theory Review 5*(3), 229-247.

Griffin, P. B. 1982 The acquisition and sharing of food among Agta foragers. *Paper presented at "The Sharing of Food: From Phylogeny to History" Conference, Homburg, Germany.*

Grove, M.; Pierce, E., & Dunbar, R. 2012. Fission-fusion and the evolution of hominin social systems. *Journal of Human Evolution 62*(2), 191-200.

Gruter, M., & Masters, R. D. 1986. Ostracism: a social and biological phenomenon. *Ethology and Sociobiology 7*(3-4),149–395.

Gurven, M. 2004. To give and to give not: the behavioral ecology of human food transfers. *Behavioral and Brain Sciences 27*(4), 543-559.

Gurven, M., & Hill, K. 2009. Why do men hunt? *Current Anthropology 50*(1), 51–73.

Gurven, M.; Hill, K.; Kaplan, H.; Hurtado, M., & Lyles, B. 2000. Food transfers among Hiwi foragers of Venezuela: Tests of reciprocity. *Human Ecology* 28, 171–218.

Haidt, J.; Koller, S., & Dias, M. 1993. Affect, culture, and morality, or is it wrong to eat your dog? *Journal of Personality and Social Psychology 65*(4), 613–628.

Haley, K., & Fessler, D. 2005. Nobody' s watching? Subtle cues affect generosity in an anonymous economic game. *Evolution and Human Behavior 26*(3), 245–56.

Hames, R. 2000. Reciprocal altruism in Yanomamo food exchange. In N. Chagnon, L. Cronk, and W. Irons (Eds.), *Human behavior and adaptation: An anthropological perspective*. New York: Aldine de Gruyter.

Hamilton, W. D. 1963. The evolution of altruistic behavior. *The American Naturalist 97*(896), 354–356.

Hamilton, W. D. 1964. The genetical evolution of social behaviour, I & II. *Journal of Theoretical Biology 7*(1), 1–52.

Hamilton, W. D. 1970. Selfish and spiteful behaviour in an evolutionary model. *Nature 228*(5277), 1218–1220.

Hamilton, W. D. 1996. *Narrow Roads of Gene Land, Vol. I: Evolution of Social Behaviour.* Oxford: W. H. Freeman.

Han, T. A. 2013. *Intention Recognition, Commitment and Their Roles in the Evolution of Cooperation.* New York: Springer.

Handfield, T.; Thrasher, J., & García, J. 2018. Green beards and signalling: Why morality is not indispensable: Commentary/Stanford: The difference between ice cream and Nazis. *Behavioral and Brain Sciences, 14.*

Hare, B., & Tomasello, M. 2004. Chimpanzees are more skilful in competitive than in cooperative cognitive tasks. *Animal Behaviour 68*(3), 571–581.

Hare, B.; Call, J.; & Tomasello, M. 2001. Do chimpanzees know what conspecifics know? *Animal Behaviour 61*(1), 139–151.

Hare, B.; Call, J.; Agnetta, B., & Tomasello, M. 2000. Chimpanzees know what conspecifics do and do not see. *Animal Behaviour 59*(4), 771–785.

Hare, B.; Melis, A.; Woods, V.; Hastings, S., & Wrangham, R. 2007. Tolerance allows bonobos to outperform chimpanzees in a cooperative task. *Current Biology 17*(7), 619-623.

Hauser, M. D. 1997. The Evolution of Communication. Cambridge, MA: MIT Press.

Hawkes, K. 1991. Showing off: Tests of an hypothesis about men's foraging goals. *Ethology and Sociobiology 12*(1), 29–54.

Hawkes, K., & Bliege Bird, R. L. 2002. Showing off, handicap signaling, and the evolution of men's work. *Evolutionary Anthropology 11*(2), 58–67.

Hawkes, K.; O'Connell, J. F., & Blurton Jones, N. G. 1991. Hunting income patterns among the Hadza: Big game, common goods, foraging goals and the evolution of the human diet. *Philosophical Transactions of the Royal Society of London B* 334, 243–251.

Hawkes, K.; O'Connell, J. F., & Blurton Jones, N. G. 2001. Hadza meat sharing. *Evolution and Human Behavior 22*(2), 113–142.

Hayek, F. A. 1982. *Law, Legislation and Liberty: A New Statement of the Liberal Principles of Justice and Political Economy*. London: Routledge & Kegan Paul.

Henrich, J. 2004. Cultural group selection, coevolutionary processes and large-scale cooperation. *Journal of Economic Behavior & Organization 53*(1), 3-35.

Henrich, J. 2015. *The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter.* Princeton University Press.

Henrich, J. 2020. *The WEIRDest People in the World: How the West Became Psychologically Peculiar and Particularly Prosperous.* New York: Farrar, Straus and Giroux.

Henrich, J., & Boyd, R. 2001. Why People Punish Defectors: Weak Conformist Transmission can Stabilize Costly Enforcement of Norms in Cooperative Dilemmas. *Journal of Theoretical Biology 208*(1), 79-89.

Henrich, J.; McElreath, R.; Barr, A., et al. 2006. Costly Punishment Across Human Societies. *Science 312*(5781)*,* 1767-1770.

Henzi, S. P., & Barrett, L. 2002. Infants as commodity in a baboon market. *Animal Behaviour 63*(5)*,* 915–921.

Hess, N. H., & Hagen, E. H. 2006. Psychological adaptations for assessing gossip veracity. *Human Nature 17*(3), 337–354.

Hirshleifer, J. 1984. On the Emotions as Guarantors of Threats and Promises. *UCLA Economics Working Papers 337*, *UCLA Department of Economics.*

Hirshleifer, J. 2003. Game-Theoretic Interpretations of Commitment. In R. Nesse (Ed.), *Evolution and the Capacity for Commitment,* pp. 77-91. New York: Russell Sage Foundation.
Holmberg, A. R. 1969. *Nomads of the Long Bow – The Siriono of Eastern Bolivia.* New York: Natural History Press.

Holmes, R. 1985. *Acts of war: Behavior of Men in Battle*. New York: Free Press.

Hopfensitz, A., & Reuben, E. 2009. The importance of emotions for the effectiveness of social punishment. *The Economic Journal 119*, 1534–59.

Hrdy, S. B. 2009. *Mothers and others: the evolutionary origins of mutual understanding.* Cambridge, MA: Harvard University Press.

Huebner, B.; James, J. L., & Hauser, M. D. 2010. The Moral-Conventional Distinction in Mature Moral Competence. *Journal of Cognition and Culture 10*, 1–26.

Hurford, J. R. 2007. Language in the Light of Evolution I: The Origins of Meaning. Oxford: Oxford University Press.

Hurford, J. R. 2011. The origins of meaning. In M. Tallerman & K. R. Gibson (Eds.), *The Oxford Handbook of Language Evolution,* pp. 370-81. Oxford: Oxford University Press.

Hyden, G. 2006. Between State and Community: Challenges to Redesigning Governance in Africa. *Working Conference on "Designing Constitutional Arrangements for Democratic Governance in Africa: Challenges and Possibilities," held at the Workshop in Political Theory and Policy Analysis, Indiana University Bloomington, March 30-31, 2006.*

Irons, W. 2003. Religion as a Hard-to-Fake Sign of Commitment. In R. Nesse (Ed.), *Evolution and the Capacity for Commitment*, pp. 292-309. New York: Russell Sage Foundation.

Jordan, J. J.; Hoffman, M.; Bloom, P., & Rand, D. 2016. Third-party punishment as a costly signal of trustworthiness. *Nature 530,* 473–476.

Jordan, J. J.; Sommers, R.; Bloom, P., & Rand, D. G. 2017. Why do we hate hypocrites? Evidence for a theory of false signaling. *Psychological Sciences 28*, 356–68.

Kaplan, H.; Gurven, M. 2005. The natural history of human food sharing and cooperation: A review and a new multi- individual approach to the negotiation of norms. In S. Bowles, R. Boyd, E. Fehr & H. Gintis (Eds.), *The Moral Sentiments and Material Interests: The Foundations of Cooperation in Economic Life,* pp. 75-115. Cambridge, MA: MIT Press.

Kaplan, H.; Hill, K.; Lancaster, J., & Hurtado, M. 2000. A theory of human life history evolution: Diet, intelligence, and longevity. *Evolutionary Anthropology 9*(4): 156 – 185.

Kaufmann, W.; Hooghiemstra, R., & Feeney, M. K. 2018. Formal institutions, informal institutions, and red tape: A comparative study. *Public Administration 96*(2), 386-403.

Kerr, N. L.; Rumble, A. C.; Park, E. S.; Ouwerkerk, J. W.; Parks, C. D,; Gallucci, M & Van Lange, P. A. M. 2009. How many bad apples does it take to spoil the whole barrel?: Social exclusion and toleration for bad apples. *Journal of Experimental Social Psychology 45*(4), 603–613.

Kingma, S. A. 2017. Direct benefits explain interspecific variation in helping behaviour among cooperatively breeding birds. *Nature Communications 8*, 1-7.

Koehler, J. J., & Gershoff, A. D. 2003. Betrayal aversion: When agents of protection become agents of harm. *Organizational Behavior and Human Decision Processes, 90*(2), 244–261.

Kohlberg, L. 1984. *The psychology of moral development: The nature and validity of moral stages.* New York: Harper and Row.

Kroodsma, D. E. 1977. Correlates of song organization among North American wrens. *The American Naturalist 111*(981), 995–1008.

LaCroix, T. 2020. Evolutionary Explanations of Simple Communication: Signalling Games and Their Models. *Journal for General Philosophy of Science 51*, 19–43.

Laidre, M. E., & Johnstone, R. A. 2013. Animal Signals. *Current Biology 18,* 829-833.

Layton, R., & O'Hara, S. 2010 Human social evolution: a comparison of hunter-gatherer and chimpanzee social organization. In R. Dunbar, C. Gamble & J. Gowlett (Eds.), *Social brain, distributed mind*, pp. 83-113. Oxford: Oxford University Press.

Layton, R.; O'Hara, S., & Bilsborough, A. 2012. Antiquity and social function of multilevel social organization among human hunter-gatherers. *International Journal of Primatology* 33(5), 215–1245.

Lewis, D. 1969. *Convention.* Oxford: Blackwell.

Liebal, K.; Behne, T.; Carpenter, M., & Tomasello, M. 2009. Infants use shared experience to interpret pointing gestures. *Developmental Science 12*(2), 264– 271.

Liszkowski, U.; Schäfer, M.; Carpenter, M., & Tomasello. M. 2009. Prelinguistic infants, but not chimpanzees, communicate about absent entities. *Psychological Science 20*(5), 654– 660.

Lombard, M., & Shea, J. J. 2021. Did Pleistocene Africans use the spearthrower-and-dart? *Evolutionary Anthropology 30*(5), 307-315.

Lorini. G. 2018. Animal Norms: An Investigation of Normativity in the Non-Human Social World. *Law, Culture and the Humanities,* 1-22.

Luce, R. D., & Raiffa, H. 1957. *Games and decisions: Introduction and critical survey.* New York: Wiley.

Luncz, L. V.; Mundry, R., & Boesch, C. 2012. Evidence for cultural differences between neighboring chimpanzee communities. *Current Biology 22*(10), 922-6.

Machery, E. 2012. Delineating the moral domain. In *The Baltic International Yearbook of Cognition, Logic and Communication: Vol. 7*, 1-14.

Malešević, S. 2021. Emotions and Warfare: The Social Dynamics of Close-Range Fighting. *Oxford Research Encyclopedia of Politics*.

Marlowe, F. W. 2005. Hunter-gatherers and human evolution. *Evolutionary Anthropology 14*(2), 54 – 67.

Massar, K. 2021. Disgust. In T. K. Shackelford & V. A. Weekes-Shackelford (Eds.), *Encyclopedia of Evolutionary Psychological Science*. Springer, Cham.

Mathew, S., & Boyd, R. Punishment sustains large-scale cooperation in pre-state warfare. *The Proceedings of the National Academy of Sciences 108*(28), 11375-11380.

Mayr, E. 1961. Cause and effect in biology. *Science 134*, 1501–1506.

McClennen, E. F. 1990. *Rationality and Dynamic Choice*. Cambridge: Cambridge University Press.

McComb, K., & Semple, S. 2005. Coevolution of Vocal Communication and Sociality in Primates. *Biology Letters 1*, 381–385.

McNeill, W. 1997. *Keeping together in time: Dance and drill in the human history.* Cambridge, MA: Harvard University Press.

Melis, A.; Hare, B., & Tomsello, M. 2006. Engineering cooperation in chimpanzees: tolerance constraints on cooperation. *Animal Behavior 72*(2), 275-286.

Migliano, A. B.; Battison, F.; Viguier, S., et al. 2020. Hunter-gatherer multilevel sociality accelerates cumulative cultural evolution. *Science Advances 6*(9), 1-7.

Mischkowski, M.; Stone, R., & Stremitzer, A. 2019. Promises, Expectations, and Social Cooperation. *The Journal of Law and Economics 62*(4).

Mohtashemi, M., & Mui, L. 2003. Evolution of indirect reciprocity by social information: the role of trust and reputation in evolution of altruism. *Journal of Theoretical Biology 223*, 523-531.

Molho, C.; Tybur, J. M.; Van Lange, P. A. M & Balliet, D. 2020. Direct and indirect punishment of norm violations in daily life. *Nature Communications 11*, 3432.

Moll, H.; Koring, C.; Carpenter, M., & Tomasello, M. 2006. Infants determine others' focus of attention by pragmatics and exclusion. *Journal of Cognition and Development 7*(3), 411– 430.

Myers, F. 1988. Burning the truck and holding the country: Property, time and the negotiation of identity among Pintupi Aborigines. In T. Ingold, D. Riches & J. Woodburn (Eds.), *Hunter-gatherers, vol. II: Property, power and ideology*. New York: Routledge.

Nakamaru, M., & Kawata, M. 2004. Evolution of rumours that discriminate lying defectors. *Evolutionary Ecology Research 6*, 261-283.

Nesse. R. 2003. Natural Selection and the Capacity for Subjective Commitment. In R. Nesse (Ed.), *Evolution and the Capacity for Commitment,* pp. 1-44. New York: Russell Sage Foundation.

Nichols, S. 2004. *Sentimental Rules: On the Natural Foundations of Moral Judgment.* Oxford: Oxford University Press.

Nichols, S., & Folds-Bennett, T. 2003. Are children moral objectivists? Children's judgments about moral and response-dependent properties. *Cognition 90*(2), B23–32.

Nichols, S.; Svetlova, M., & Brownell, C. 2009. The role of social understanding and empathic disposition in young children's responsiveness to distress in parents and peers. *Cognition, Brain, Behaviour 4,* 449-478.

Nikiforakis, N. 2008. Punishment and counter-punishment in public good games: Can we really govern ourselves? *Journal of Public Economics 92*(1-2), 91–112.

Nowak M. A., & Sigmund K. 1998. Evolution of indirect reciprocity by image scoring. *Nature 393*, 573-577.

Nowak, M. A., & Sigmund, K. 2005. Evolution of indirect reciprocity. *Nature 437*, 1291–1298.

Nucci, L. 2001. *Education in the Moral Domain*. Cambridge, MA: Cambridge University Press.

Nucci, L., & Turiel, E. 1993. God's word, religious rules, and their relation to Christian and Jewish children's concepts of morality. *Child Development 64*(5), 1475-1491.

Nucci, L.; Turiel, E., & Encarnacion-Gawrych, G. 1983. Children's social interactions and social concepts in the Virgin Islands. *Journal of Cross-Cultural Psychology 14*(4), 469-487.

Oka, R., & Kusimba, C. 2008. The archaeology of trading systems part 1: towards a new trade synthesis. *J. Archaeol. Res. 16,* 339-395.

Ostner, J. 2018. Primate Social Cognition: Evidence from Primate Field Studies. In L. D. Di Paolo et al. (eds), *Evolution of Primate Social Cognition*. Interdisciplinary Evolution Research 5, 97-106. New York: Springer.

Ozono, H.; Jin, N.; Watabe, M., & Shimizu, K. 2016. Solving the second-order free rider problem in a public goods game: An experiment using a leader support system. *Scientific Reports 6:38349,* 1-8.

Panchanathan K., & Boyd R. 2003. A tale of two defectors: The importance of standing for evolution of indirect reciprocity. *Journal of Theoretical Biology 224*, 115-126.

Pejovich, S. 1999. The effects of the interaction of formal and informal institutions on social stability and economic development. *Journal of Markets and Morality 2*, 164– 181.

Peters, K., & Kashima, Y. 2013. Gossiping as moral social action. In J. P. Forgas; O. Vincze & J. Laszlo (eds), *Social cognition and communication*, 187–202. New York: Psychology Press.

Piaget, J. [1932] (1965). *The psychology of moral development: The nature and validity of moral stages.* Translated by M. Gabain. New York: Free Press.

Pickering, T. R. 2013. *Rough and Tumble: Aggression, Hunting, and Human Evolution*. Berkeley, CA: University of California Press.

Pleyer, M., & Hartmann, S. 2019. Constructing a consensus on language evolution? Convergences and differences between biolinguistic and usage-based approaches. *Frontiers in Psychology 10*, 2537.

Pollard, K. A., & Blumstein, D. T. 2012. Evolving communicative complexity: insights from rodents and beyond. *Philosophical Transactions of the Royal Society B: Biological Sciences 367*(1597), 1869–1878.

Pollock, G., & Dugatkin, L. A. 1992. Reciprocity and the emergence of reputation. *Journal of Theoretical Biology 159*(1), 25-37.

Powell, C. A. J., & Smith, R. H. 2013. Schadenfreude caused by the exposure of hypocrisy in others. *Self and Identity 12*(4), 413–431.

Powers, S. T., & Lehmann, L. 2013. The co-evolution of social institutions, demography, and large-scale human cooperation. *Ecology Letters 16,* 1356-1364.

Powers, S. T.; van Schaik, C. P., & Lehmann, L. 2016. How institutions shaped the last major evolutionary transition to large-scale human societies. *Philosophical Transactions of the Royal Society B: Biological Sciences 371,* 1-10.

Puri, R. K. 2005. Deadly dances in the Bornean rainforest: Hunting knowledge of the Penan Benalui. Leiden, KITLV.

Rapoport, A., & Chammah, M. 1965. *Prisoner's Dilemma: A Study in Conflict and Cooperation*. Ann Arbor: University of Michigan Press.

Richerson, P., & Boyd, R. 1998. The Evolution of Human Ultra-sociality. In I. Eibl-Eibisfeldt & F. Salter (Eds.), *Ideology, Warfare, and Indoctrinability*. New York: Berghahn Books.

Robson, A. 1990. Efficiency in evolutionary games: Darwin, Nash, and the secret handshake. *Journal of Theoretical Biology 144*, 379–396.

Rogers, H. R. 2017. Memory Retention Rates of Gossip-Related Information. *Honors Theses.* 373. https://egrove.olemiss.edu/hon_thesis/373.

Rozin, P., Haidt, J., & McCauley, C. 2000. Disgust. In M. Lewis & S. M. Haviland-Jones (Eds.), *Handbook of emotions* (2nd ed., pp. 637–653). New York: Guilford Press.

Rubenstein, D. R., & Lovette, I. J. 2007. Temporal environmental variability drives the evolution of cooperative breeding in birds. *Current Biology 17*(16), 1414-9.

Rudert, S. C.; Ruf, S., & Greifeneder, R. 2020. Whom to punish? How observers sanction norm-violating behavior in ostracism situations. *European Journal of Social Psychology 50*(2), 376–391.

Sachs, J. L.; Mueller, U. G.; Wilcox, T. P., & Bull, J. J. 2004. The evolution of cooperation. *Quarterly Review of Biology 79*(2), 135–160.

Sahlins, M. D. 1963. Poor Man, Rich Man, Big-Man, Chief: Political Types in Melanesia and Polynesia. *Comparative Studies in Society and History 5*(2), 285-303.

Santos, F. C.; Pacheco, J. M., & Skyrms, B. 2011. Co-evolution of pre-play signaling and cooperation. *Journal of Theoretical Biology 274,* 30–35.

Sasaki, T & Uchida S. 2013. The evolution of cooperation by social exclusion. *Proceedings of the Royal Society B: Biological Sciences 280: 20122498,* 1-7.

Schelling, T. C. 1960. *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.

Schelling, T. C. 2003. Commitment: Deliberate Versus Involuntary. In R. Nesse (Ed.), *Evolution and the Capacity for Commitment,* pp. 48-56. New York: Russell Sage Foundation.

Schlingloff, L., & Moore, R. 2018. Do Chimpanzees Conform to Social Norms. In K. Andrews & J. Beck (Eds.), *The Routledge Handbook Of Philosophy Of Animal Minds.* Oxford: Routledge.

Schmidt, M. F. H.; Gonzalez-Cabrera, I., & Tomasello, M. 2017. Children's developing metaethical judgments. *Journal of Experimental Child Psychology 164*, 163-177.

Schoenmakers, S.; Hilbe, C.; Blasius, B., & Traulsen, A. 2014. Sanctioning as honest signals – The evolution of pool punishment by sanctioning institutions. *Journal of Theoretical Biology 356,* 36-46.

Searle, J. R. 2003. Social Ontology and Political Power. In F. F. Schmitt (Ed.), *Socializing Metaphysics: The Nature of Social Reality.* Oxford: Rowman & Littlefield.

Selman, R. 1980. *The growth of interpersonal understanding*. New York: Academic Press.

Seyfarth, R. M., & Cheney, D. L. 2012. Animal Cognition: Chimpanzee Alarm Calls Depend On What Others Know. *Current Biology 22*(2), 51–52.

Shea, J. 2009. The impact of projectile weaponry on Late Pleistocene hominin evolution. In J. Hublin & M. P. Richards (Eds.), *The Evolution of Hominid Diets,* pp. 187 – 198. Berlin: Springer Science.

Skitka, L. J.; Bauman, C. W., & Sargis, E. G. 2005. Moral conviction: Another contributor to attitude strength or something more? *Journal of Personality and Social Psychology 8*, 895–917.

Skyrms, B. 1996. *The Evolution of the Social Contract*. Cambridge: Cambridge University Press.

Skyrms, B. 2002. Signals, evolution, and the explanatory power of transient information. *Philosophy of Science 69*, 407–428.

Skyrms, B. 2003. *The Stag Hunt and the Evolution of Social Structure.* Cambridge: Cambridge University Press.

Skyrms, B. 2004. The Stag Hunt and the Evolution of Social structure. Cambridge: Cambridge University Press.

Skyrms, B. 2010. Signals: Evolution, Learning, & Information. Oxford: Oxford University Press.

Skyrms, B., & Barrett, J. 2018. Propositional Content in Signals. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences 74*, 34-39.

Smetana, J., & Braeges, J. 1990. The Development of Toddlers' Moral and Conventional Judgments. *Merrill-Palmer Quarterly 36*(3), 329-346.

Smith, E. A. 2003. Human cooperation: Perspectives from behavioral ecology. In P. Hammerstein (ed), *The Genetic and Cultural Evolution of Cooperation*, 401−427. Cambridge, MA: MIT Press.

Smith, E. A. 2010. Communication and collective action: language and the evolution of human cooperation. *Evolution and Human Behavior 31*, 231–245.

Smith, E. A.; Bliege Bird, R., & Bird, D. 2003. The benefits of costly signaling: Meriam turtle hunters. *Behavioral Ecology 14*(1), 116–126.

Sober, E., & Wilson, D. S. 1998. *Unto others: The evolution and psychology of unselfish behavior.* Cambridge, MA: Harvard University Press.

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

Sperber, D., & Baumard, N. 2012. Moral Reputation: An Evolutionary and Cognitive Perspective. *Mind and Language 27*(5), 495-518.

Stanford, P. K. 2018a. The Difference Between Ice Cream and Nazis. *Behavioral and Brain Sciences 14*, 1-13.

Stanford, P. K. 2018b. Moral Externalization and Normativity: The Errors of Our Ways. *Behavioral and Brain Sciences 14*, 34-45.

Sterelny, K. 2012. *The Evolved Apprentice: How Evolution made Humans Unique.* Cambridge, MA: MIT Press.

Sterelny, K. 2019. Religion: costs, signals, and the Neolithic transition. *Religion, Brain & Behavior 10*(3), 303-320.

Sterelny, K. 2021a. *The Pleistocene Social Contact: Culture and Cooperation in Human Evolution.* Oxford: Oxford University Press.

Sterelny, K. 2021b. The Origins of Multi-level Society. *Topoi 40,* 207-220.

Sterelny, K. 2021c. How equality slipped away. *Aeon*, 10 June 2021.

Sterelny, K., & Planer, R. J. 2021. *From Signal to Symbol: The Evolution of Language.* Cambridge, MA: MIT Press.

Stiner, M. C. 2001. Thirty years on: The “Broad Spectrum Revolution” and Paleolithic demography. *The Proceedings of the National Academy of Sciences 98*(13), 6993-6996.

Stiner, M. C.; R. Barkai & A. Gopher. 2009. Cooperative hunting and meat sharing 400– 200 kya at Qesem Cave, Israel. *The Proceedings of the National Academy of Sciences 106*(32), 13207– 13212.

Stroud, L. R.; Tanofsky-Kraff, M.; Wilfley, D. E., & Salovey, P. 2000. The Yale Interpersonal Stressor (YIPS): affective, physiological, and behavioral responses to a novel interpersonal rejection paradigm. *Annals of Behavioral Medicine 22*(3), 204–13.

Stuart-Fox. M. 2022. Major Transitions in Human Evolutionary History. *World-Futures: The Journal of New Paradigm Research 79*(1), 29-68.

Sylwester, K., & Roberts, G. Cooperators benefit through reputation-based partner choice in economic games. *Biology Letters 6*(5), 659-662.

Tallerman, M., & Gibson, K. R. 2011. *The Oxford Handbook of Language Evolution*. Oxford: Oxford University Press.

Tinbergen, N. 1968. On war and peace in animals and man. *Science 160*, 1411–18.

Tomasello, M. 2008. *Origins of Human Communication.* Cambridge, MA: MIT Press.

Tomasello, M. 2009. *Why We Cooperate.* Cambridge, MA: MIT Press.

Tomasello, M. 2014. *A Natural History of Human Thinking.* Cambridge, MA: Harvard University Press.

Tomasello, M. 2016. *Natural History of Human Morality.* Cambridge, MA: Harvard University Press.

Tomasello, M., & Carpenter, M. 2007. Shared intentionality. *Developmental Science 10* (1), 121–125.

Tomasello, M.; Carpenter, M.; Call, J.; Behne, T., & Moll, H. 2005. Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences 28*(5), 675–691.

Tomasello, M.; Melis, A. P.; Claudio, T.; Wyman, E., & Herrmann, E. 2012. Two Key Steps in the Evolution of Human Cooperation: The Interdependence Hypothesis. *Current Anthropology 53*(6), 673-692.

Trivers, R. 1971. The Evolution of Reciprocal Altruism. *Quarterly Review of Biology 46*(1), 35-57.

Turchin, P.; Currie, T. E.; Turner, E. A., & Gavrilets, S. 2013. War, space, and the evolution of Old World complex societies. *Proc Natl Acad Sci 110*(41), 16384-9.

Turiel, E. 1979. Distinct conceptual and developmental domains: Social convention and morality. In H. Howe & C. Keasey (eds), *Nebraska Symposium on Motivation, 1977: Social Cognitive Development*, 77-116. Lincoln, NE: University of Nebraska Press.

Turiel, E., & Nucci, L. 1978. Social interactions and the development of social concepts in preschool children. *Child Development 49*(2), 400-407.

Turiel, E.; Killen, M., & Helwig, C. 1987. Morality: Its structure, functions, and vagaries. In J. Kagan & S. Lamb (Eds.), *The Emergence of Morality in Young Children*, pp. 155–244. Chicago: University of Chicago Press.

Tybur, J. M., Lieberman, D., & Griskevicius, V. 2009. Microbes, mating, and morality: Individual differences in three functional domains of disgust. *Journal of Personality and Social Psychology 97*, 103–122.

Vanberg, C. 2008. Why Do People Keep Their Promises? An Experimental Test of Two Explanations. *Econometrica 76*(6), 1467-1480.

Vasquez, K.; Keltner, D.; Ebenbach, D. H., & Banaszynski, T. L. 2001. Cultural variation and similarity in moral rhetorics: Voices from the Philippines and the United States. *Journal of Cross-Cultural Psychology 32*(1), 93–120.

Warneken, F. 2013. Young children proactively remedy unnoticed accidents. *Cognition 126*(1), 101-108.

Warneken, F., & Tomasello, M. 2006. Altruistic helping in human infants and young chimpanzees. *Science 311*(5765), 1301-1303.

Warneken, F., & Tomasello, M. 2007. Helping and cooperation at 14 months of age. *Infancy 11*(3), 271-294.

Warneken, F., & Tomasello, M. 2009. Varieties of altruism in children and chimpanzees. *Trends in Cognitive Sciences 13*(9), 397–402.

Warneken, F., & Tomasello, M. 2013. The emergence of contingent reciprocity in young children. *Journal of Experimental Child Psychology 116*(2), 338-350.

Watkins, T. 2010. New Light on the Neolithic Revolution in South-West Asia. *Antiquity 84,* 621-634.

Watts, D. P., & Mitani, J. C. 2001. Boundary Patrols and Intergroup Encounters in Wild Chimpanzees. *Behaviour 138*(3)*,* 299-327.

Wellman, H.; Harris, P; Banerjee, M., & Sinclair, A. 1995. Early understanding of emotion: Evidence from natural language. *Cognition and Emotion 9*(2-3), 117–79.

West, S. A.; Griffin, A. S.; Gardner, A., & Diggle, S. P. 2006. Social evolution theory for microbes. *Nature Reviews, Microbiology 4*, 597–607.

West, S.; Griffin, A. S., & Gardner, A. 2007. Social Semantics: Altruism, Cooperation, Mutualism, Strong Reciprocity and Group Selection. *Journal of Evolutionary Biology 20*(2), 415–432.

Wilkinson, G. S. 1988. Reciprocal altruism in bats and other mammals. *Ethology and Sociobology 9*(2-4), 85-100.

Wilkinson, G. S. 2003. Social and vocal complexity in bats. In F. B. M. De Waal & P. L. Tyack (eds), *Animal social complexity: intelligence, culture, and individualized societies*. Cambridge, MA: Harvard University Press.

Williams, K. D. 2007. Ostracism. *Annual Review of Psychology 58*, 425–452.

Wilson, D. S. 1975. A Theory of Group Selection. *Proceedings of the National Academy of Sciences 72*(1), 143-146.

Wilson, D. S. 1977. Structured Demes and the Evolution of Group-Advantageous Traits. *American Naturalist 111*(977), 157-1 85.

Wilson, D. S. 1998. Hunting, Sharing, and Multilevel Selection: The Tolerated-Theft Model Revisited. *Current Anthropology 39*(1), 73-97.

Wilson, D. S., & Colwell, R. K. 1981. Evolution of Sex Ratio in Structured Demes. *Evolution 35*(5), 882-897.

Wilson, D. S., & Sober, E. 1994. Reintroducing group selection to the human behavioral sciences. *Behavioral and Brain Sciences 17*(4), 585–654.

Wiseman, T., & Yilankaya, O. 1999. Cooperation, Secret Handshakes, and Imitation in the Prisoner's Dilemma. *Northwestern University CMS-EMS Discussion Paper 1248*.

Wrangham, R. W. 1999. Evolution of Coalitionary Killing. *Yearbook of Physical Anthropology 42*, 1–30

Wubs, M.; Bshary, R., & Lehmann, L. 2016. Coevolution between positive reciprocity, punishment, and partner switching in repeated interactions. *Proceedings of the Royal Society B: Biological Sciences 283*(1832), 1-8.

Zhang, B.; Li, C.; De Silva, H.; Bednarik, P., & Sigmund, K. 2014. The evolution of sanctioning institutions: an experimental approach to the social contract. *Experimental Economics 17,* 285-303.

Ziker, J., & Schnegg, M. 2005. Food sharing at meals: Kin- ship, reciprocity, and clustering in the Taimyr Autonomous Region, northern Russia. *Human Nature* 16, 178–210.

Zuberbühler, K. 2011. Cooperative breeding and the evolution of vocal flexibility. In M. Tallerman & K. R. Gibson (Eds.), *The Oxford Handbook of Language Evolution,* pp. 71-81. Oxford: Oxford University Press.