When it pays to punish in the evolution of honesty and cooperation

Hannah Rubin*

Abstract

In explaining the emergence of conventions surrounding human cooperation and helping of those in need, it seems as though honest communication of need is an essential part of the story. While previous results indicate that punishment promotes cooperation, this paper will argue that the story is more complicated. Namely, whether punishment promotes cooperation depends on what you punish. Punishment of those who lie about their need for a resource may instead impede cooperation, as the attempts to deceive that arise in cooperative endeavors may be too costly to make cooperation worthwhile.

1 Introduction

It has long been argued that conventions are at the basis of our concept of justice; mutual expectations of how to act allow people to coordinate on following certain practices, to the benefit of all [Hume, 2000b,a]. There are various ways to spell out just how conventions underlie our ideas of justice, but we can say that rational agents accept certain conventions as just, and so allow them to regulate the actions they choose in pursuit of their own self-interest, when those conventions lead to mutual advantage in cases where there is some conflict of interest. But how do these conventions emerge? Many turn to evolutionary game theory for an explanation.

There are countless aspects of human conventions one may choose to focus in on and explain; two that will be important here are the role of punishment for failure to cooperate and the emergence of honest communication. Regarding punishment: it is a common observation that punishment can support social coordination and cooperation [Boyd and Richerson, 1992, Boyd et al., 2003, Fowler, 2005]. This is often discussed in relation to altruistic behavior, where one person sacrifices something to help another [Gintis, 2000, Henrich and Boyd, 2001, Bowles and Gintis, 2004], though Vanderschraaf [2016, 2018] has recently extended this reasoning to cases of cooperation captured by a stag hunt. Following Lewis [1969], much of our understanding of human communication involves

*I would like to thank Justin Bruner and two anonymous reviews for helpful comments, as well as audiences at Oberlin College and the Varieties of Information workshop on Deception.

conventions as well. Meaning can be found when there is common expectation of following one of many equilibria in a type of coordination game called a signaling game, and, furthermore, this meaning can emerge spontaneously [Skyrms, 2010].

In order to understand conventions surrounding human cooperation and helping of those in need, it seems as though these two aspects must be brought together; honest communication of need is part of a convention that produces joint benefits. More generally, cooperative interactions are often complicated, and our understanding of what factors lead to mutually beneficial conventions may change when we take into account this complexity [Zollman, 2008, Bednar and Page, 2007, Wagner, 2012]. There may be multiple types of coordination needed in order to succeed. If one of these types of coordination is not successful, this means the cooperative endeavor fails overall. For example, there is no benefit to honestly communicating whether you are in need if there is no expectation that those around you will help those in need. There are many stages at which the cooperative endeavor may be prevented if an appropriate convention has not been established.

Section 2 will set the stage for investigating the role punishment can play in this context by discussing some previous results surrounding punishment and its effects on human altruism and cooperation. Section 3 will then look at how including the evolution of honest communication, and the possibility of punishing lying, in this picture affects the conclusions we draw about the possibility of certain conventions emerging. We will see that punishment can be important to sustaining honest communication where there is some conflict of interest between those who are communicating. However, the opportunity to punish lying (somewhat surprisingly) decreases the chances for cooperation to evolve. That is, when there are multiple stages in a cooperative endeavor (e.g. honest communication preceding helping of those in need), punishment at one stage can make it harder for the cooperative endeavor as a whole to succeed. Section 4 concludes with a general lesson for studying the evolution of conventions.

2 Punishment, cooperation, and donation

Costly punishment is commonly talked about in explanations of human cooperation and altruism. Cases where people (or other organisms) help others at a cost to themselves, or with no apparent benefit to themselves, are traditionally a puzzle for evolutionary theory – why would that behavior evolve when it seems to lower the success of that organism compared to others that do not exhibit the behavior? This sort of puzzling situation is often represented by the Prisoner's dilemma (table 1a) in game theoretic models. In this game, cooperation always has a lower payoff than defection, but cooperation increases the payoff of the other player. One scenario that the prisoners' dilemma usefully captures is the altruistic donation of resources to someone in need, where a cooperator pays a cost to donate something of value to their interactive partner, while a defector

does not donate.[1] For instance, the payoffs in table 1a would arise if individuals start off with a 'baseline' payoff of 1 and cooperators pay a cost of 1 to donate something of value 2 to their interactive partner.

	C	D
C	2,2	0,3
D	3,0	1,1

(a)

	C	P	D
C	2,2	2,2	0,3
P	2,2	2,2	-1,1
D	3,0	1,-1	1,1

(b)

Table 1: Example Prisoners' Dilemma and Augmented Prisoners' Dilemma. (a) Prisoners' Dilemma, with possible strategies to cooperate, C, or defect, D. (b) Augmented Prisoners' Dilemma, where players have the additional option to cooperate and punish defectors, P.

There are many different explanations that have been developed for how altruistic/cooperative behavior could evolve, but when it comes to human behavior it is common to talk about social costs of not cooperating. Many argue that the presence of so-called *strong reciprocators*, who both cooperate and punish non-cooperators, makes it more likely for cooperation to evolve and can stabilize cooperative norms in a population [Gintis, 2000, Henrich and Boyd, 2001, Bowles and Gintis, 2004]. Table 1b shows an example of a prisoners' dilemma augmented to include this possible punishing strategy, where a strong reciprocator can pay a cost of 1 to inflict a punishment of 2 on a defector. This yields a new game, which is no longer a prisoners' dilemma but which we can call an *augmented prisoners' dilemma*, to match Vanderschraaf [2016]'s terminology in describing the augmented stag hunt (see below). With strong reciprocity, cooperative behavior is not so puzzling – with enough people who are willing to punish non-cooperative behavior, cooperation becomes appealing because it allows people to avoid being punished.

However, this seems to push the puzzle elsewhere, because these strong reciprocators are generally assumed to have to pay a cost to punish non-cooperators – why would *that* sort of behavior evolve when it seems to lower the success compared to others that do not exhibit the behavior? In this game, the punishing strong reciprocator strategy is *weakly dominated*; regardless of what the other player chooses, cooperators always do at least as well and sometimes do better. While it seems counter-intuitive that a rational agent would adopt such a strategy, many have shown that that weakly dominated strategies can evolve and persist in a population [e.g., Gale et al., 1995, Samuelson, 2002, Skyrms, 2014]. In the context of strong reciprocity, there are various ways to show that strong reciprocity can evolve, commonly by appealing to group selection and arguing, for instance, that groups with disproportionately many strong reciprocators are better able to survive due to people within that group helping each

[1] This is considered 'altruistic' in the sense that the cooperator lowers their own payoff to increase someone else's payoff. Labelling this behavior altruistic does not necessarily imply anything about the intentions of the actor.

other out and ensuring each other's survival [see, e.g. Boyd et al., 2003, Bowles and Gintis, 2011].[2]

While punishment is commonly talked about in relation to altruistic behavior, Peter Vanderschraaf [2016, 2018] looks at punishment in a different scenario: a stag hunt, where cooperation is mutually beneficial (table 2a).[3] In the traditional story, two players can either cooperate to hunt a large game (stag) or go it alone and hunt smaller prey (hare). When both players cooperate to hunt a stag, their joint payoffs are better than when either or both defect and hunt hare alone. However, hunting stag is risky – if the other person does not cooperate, your hunt will not be successful and you will end up empty handed. More specifically, hunting hare is *risk dominant*, i.e. it has a larger basin of attraction, meaning evolution more often leads to hare hunting than stag hunting. By contrast, hunting stag is *Pareto efficient*, meaning no one could do better in another situation without making someone else worse off. In other words, the cooperative equilibrium where both hunt stag is more desirable, but the equilibrium where both hunt hare is more likely to emerge.

	C	D
C	3,3	0,2
D	2,0	2,2

(a)

	C	P	D
C	3,3	3,3	0,2
P	3,3	3,3	-1,0
D	2,0	0,-1	2,2

(b)

Table 2: Example Stag Hunt and Augmented Stag Hunt. (a) Stag Hunt, with possible strategies to cooperate to hunt stag, C, or defect to hunt hare alone, D. (b) Augmented Stag Hunt, with possible additional strategy to punish provocative defection.

Vanderschraaf considers how augmenting the stag hunt with a punishing strategy affects how likely it is that cooperative stag hunting will evolve [Vanderschraaf, 2016, 2018]. In particular, he considers a strategy that punishes 'provocative' defection, which occurs when one player hunts hare while their partner hunts stag. This strategy is akin to strong reciprocity, where a player will cooperate and then punish those who do not cooperate. Table 2b shows an example of this game, where the punisher pays 1 to inflict a penalty of 2 on the provocative defector. Again, in this game, the punishing strategy is weakly dominated; cooperation is always at least as good and is sometimes better. However, the possibility of punishment increases the basin of attraction for stag hunting, and, in certain cases, can make stag hunting more likely to evolve than hare hunting. The likely outcome is a population that is composed of a mix of non-punishing stag hunters and stag hunters who also punish. So, we see that "in a weakly dominated strategy there is strength": though punishing is

[2]See, e.g., Boyd and Richerson [1988], Richerson and Boyd [2008], Bowles and Gintis [2011] for arguments that group selection is an important evolutionary force in human populations.

[3]Though mutual cooperation in a prisoners' dilemma is better than mutual defection, it is not the case that an individual's cooperation is mutually beneficial – it can only ever benefit the other person. Thanks to an anonymous reviewer for pointing out this potential confusion.

	C	D
C	2.5,2.5	0,3
D	3,0	1,1

(a)

	TFT	D
TFT	5,5	1,4
D	4,1	2,2

(b)

Table 3: Example of (a) a prisoners' dilemma, and (b) a stag hunt that arises as the result of repetition of a prisoners' dilemma.

irrational in a sense (it is weakly dominated), it allows the evolution of cooperation in the stag hunt and can persist in a population where everyone cooperates [Vanderschraaf, 2016].

Cooperative hunting is not the only interpretation of the stag hunt. It is also used to model a *repeated* prisoners' dilemma [Skyrms, 2004]. In a repeated prisoners' dilemma, there are a greater number of strategies than those considered in the original, one-shot, version of the game. Instead of just cooperating or defecting, players can condition their action in one round on their partner's behavior in the previous round. A popular strategy to consider is *tit-for-tat* (TFT), which starts out cooperating then copies whatever the other player did in the previous round. This is often called a 'reciprocal altruist' strategy, since TFT players will cooperate with others who cooperate with them. If we consider TFT and defect as possible strategies, a prisoner's dilemma can be transformed into a stag hunt through repetition of the game. Table 3b shows an example of this, where players play the prisoners' dilemma in table 3a with each other twice.

These two interpretations of the stag hunt capture different aspects of human behavior, and both are important to our understanding of conventions coordinating our behavior. I will focus on this second interpretation, which is more closely related to the original scenario in which strong reciprocity is discussed (as section 3.1 elaborates). However, there is another reason to focus on the second interpretation: it will allow me to incorporate another aspect of cooperation that is not generally discussed in these cases, which is the evolution of honest communication. In order for there to be reciprocal donation of resources to those in need, there needs to be some communication between players about whether or not they are in need before donation occurs.

So, this paper will talk about punishment as relevant to another important aspect of conventions, surrounding honest communication. When there are common expectations of how and when to communicate one's condition of need, this can coordinate actions to mutual benefit of all. We might think of deception, or lying, as a sort of failure to cooperate, even though it is not usually what people consider when aiming to explain altruism and cooperative helping in humans. As we will see, there are some similar lessons we can draw about punishment in this context (e.g. it can allow honest communication to evolve), but there are also different upshots as to how punishment affects evolution of cooperation. In all, when there are multiple stages (e.g. honest communication followed by reciprocal donation) in a cooperative endeavor, punishment at one

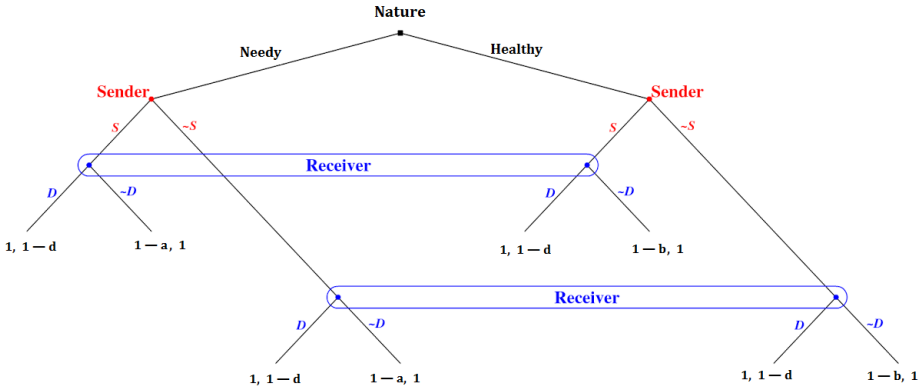


Figure 1: Extensive form of the Sir Philip Sidney game.

stage can make the cooperative endeavor as a whole less likely to succeed.

3 Models overview

The Sir Philip Sidney game, shown in figure 1, is a model the evolution of honest communication of need [Maynard Smith, 1991]. This game is most commonly used in the animal communications literature to investigate biological evolution of communication between relatives, though it is used in explanations of learned human communication norms as well. (More on this below.) Unlike many of the games discussed by Lewis [1969], in this game there is partial conflict of interest between two parties: unless you are currently in need of a donation, you have incentive to lie and pretend that you are in need, while the other party prefers that you honestly communicate.

In this game, there are two players: a sender and a receiver. The sender can be in one of two states. The sender is *needy* with probability m or *healthy* with probability $1 - m$. Which state the sender is in is known by the sender but not the receiver. The sender has two options to try to communicate with the receiver: send a signal or not. The receiver then observes whether the signal was sent and decides whether or not to donate a resource to the sender. Donation is costly as the receiver is giving up something of value. If the receiver donates, their payoff drops from 1 to $1 - d$ (where $d > 0$). A healthy individual will benefit from a donation, though not as much as someone in need. In the absence of a donation, needy and healthy senders obtain a payoff of $1 - a$ and $1 - b$, respectively (where $a > b$). If the receiver donates, then the sender obtains a payoff of 1 regardless of what state they are in.

Traditionally, it is assumed there is some cost to produce a signal, but, here it will be assumed that signals are free to send. As will be explained below, instead of assuming production costs, which may be unrealistic for humans stating they are in need of a resource, costs will be added later in the form of social costs (i.e.

punishment) for signaling when healthy. There are a few other models of the Sir Philip Sidney game which include social costs. Some are intended to capture biological evolution of interactions among relatives, e.g. offspring begging for resources from a parent [Catteeuw et al., 2014, Rich and Zollman, 2016]. As such, they assume that sender and receiver are genetically related, which affects the evolution of communication.[4]

Boyd and Mathew [2015] show how third party punishment can stabilize honest communication in the Sir Philip Sidney game among unrelated individuals, when the game is repeated and individuals trade off between the sender and receiver roles for the length of the interaction. While third party punishment may have been important in many evolutionary scenarios [Fehr and Fischbacher, 2004, Mathew et al., 2013, Chavez and Bicchieri, 2013] the third party monitoring and reputation tracking assumed in this model take us quite far from the cases of interest arising from Vanderschraaf [2016], who looks at random interactions among individuals in a population that is too large for reputation tracking to plausibly be effective.[5] As such, this paper will likewise consider cases where punishment is executed directly by the affected party.

There are, of course, many ways one might incorporate punishment into this game. In addition to third party monitoring, one might punish a sender who signals need more often than one expects an individual to be needy (similar to Rich and Zollman [2016]) or punish whenever there is any mismatch between state and signal, including being too timid to signal need when you are in fact in need (similar to Catteeuw et al. [2014]). One might also think that a plausible punishment for lying would be to refuse to donate to that individual in the future.[6] This is a type of punishment similar to that which TFT and similar strategies inflict on those who defect: punishment by withholding the future benefits of cooperation. After detecting a lie, a punisher could switch to always defecting, refuse to donate in the next round, or something along those lines. Investigating the effects of this possible punishing strategy could be an interesting avenue for future work, but, following the literature discussed in section 2, here we are interested in the effects of costly, or altruistic, punishment, where the punisher pays a cost to inflict a penalty when a healthy sender claims to be needy.[7]

Note that there are similarities between the Sir Philip Sidney game and the prisoners' dilemma described above. Receivers can choose whether to donate altruistically, paying a cost to increase the payoff of the sender. This means that repetition of the game can transform it into a game where altruistic donation is a possible evolutionary outcome. Section 3.1 will explain how, if we assume

[4]However, see [Bruner and Rubin, 2020] for concerns about the way relatedness is incorporated in these sorts of models.

[5]In other words, indirect reciprocity models might be more appropriate to a stage in human evolution where individuals interacted in small groups, whereas a model with a larger population may be more appropriate for studying norms emerging at a later stage in evolutionary history.

[6]Thanks to Justin Bruner and Nick Shea for this suggestion.

[7]See Vanderschraaf [2016, p. 47-8] for a description of how punishment in this context can be considered 'altruistic'.

there is honest communication of need, repetition can transform the Sir Philip Sidney game into a stag hunt in much the same way as it does for a prisoners' dilemma. Of course, this ignores communication aspect of the game, where there may be incentive to lie and exaggerate your need in order to acquire more donations than you would with honest signaling. Starting with the simple case will allow us to see how the potential to lie, and then the potential to be punished for a lie, impact conclusions regarding the likelihood of the evolution of cooperation. These possibilities will be explored in section 3.2. Finally, section 3.3 will compare the effects of punishing lying with the effects of punishing defection to see whether punishment of defection leads to greater overall success of the cooperative endeavor, including both honesty and donation.

As there are a number of decision points and parameters in this game, there will be many simplifying assumptions necessary to get a handle on what evolution is like. For instance, we will assume that $b = d = .1$ throughout for simplicity. In other words, a person giving up resource would have benefited just as much as a healthy sender, and this benefit is fairly small relative to the benefit it provides a needy person, a, which will be allowed to vary from .2 to .4.[8] Other simplifying assumptions will be noted as they become relevant below.

3.1 Repetition with honesty

Let us start with the simplest case and assume both that there is no punishment and that honest communication of need is guaranteed. In this case, repetition can turn the Sir Philip Sidney game into a stag hunt, much like it can turn the prisoners' dilemma into a stag hunt, as discussed in section 2. Let us assume that two players interact repeatedly, and alternate between sender and receiver roles in each interaction, as in Boyd and Mathew [2015]. There are two strategies to consider in this simplified Sir Philip Sidney game: a TFT strategy that donates to a needy sender then copies the strategy of the other player, and a Defect strategy that never donates.[9] After each round, there is some chance that the game will be repeated, and the two players will interact again. Depending on this probability of repetition, there will be some expected number of repetitions of the game between the two players.[10] Table 4 summarizes the payoffs for these two strategies when then game is repeated T times in expectation.

[8] There is nothing special about these values; similar results can be obtained as long as b and d are smaller than a.

[9] It is assumed that a TFT sender can tell the type of receiver regardless of whether they were needy that round, i.e. they perceive a provocative defection, where the receiver did not intend to donate while the TFT did intend to donate, whether or not the TFT player was actually in need of donation that round. In some situations this assumption may be plausible, e.g. when a sender can notice that a receiver is ignoring any attempt at communication of need, while in other cases it may be more plausible to assume that a sender only recognizes a provocative defection if they fail to get a donation when in need. The assumption made here greatly simplifies analysis, but future work may explore alternatives.

[10] If the probability the game continues is w, then there are $T = \frac{1}{1-w}$ repetitions of the game in expectation. For instance if $w = \frac{2}{3}$, then players are expected to play the game three times when they interact.

	TFT	Defect
TFT	$.95T$, $.95T$	$(.95 - .5ma)T + .05m(T-1)$, $(.95 + .05m)T - .5ma(T-1)$
Defect	$(.95 + .05m)T - .5ma(T-1)$, $(.95 - .5ma)T + .05m(T-1)$	$(.95 + .05m - .5ma)T$, $(.95 + .05m - .5ma)T$

Table 4: repeated Sir Philip Sidney game, assuming honest communication and no punishment.

Two TFT players will always donate the resource to each other, meaning half the time they get a payoff of 1 (either they are needy and are given the resource, or they are a receiver not donating to a healthy individual) and half the time they get a payoff of .9 (either they are donating to a needy individual or are a healthy individual not receiving a donation). If we compare these payoffs to those for two Defectors, the defectors gain a bit when they are a receiver not donating to a needy sender (specifically, they gain .1 with probability $.5m$) but they lose when they are needy and not receiving a donation (they lose a with probability $.5m$). When a TFT player interacts with a defector, in the first round, the defector has a chance to gain from receiving a donation when needy ($.5m$ chance of receiving a) compared to the TFT player, while the TFT player has one extra round of possibly donating to the Defector (losing .1 with probability $.5m$).

Let $p_{11} = .95T$, $p_{12} = (.95 - .5ma)T + .05m(T-1)$, $p_{21} = (.95 + .05m)T - .5ma(T-1)$, and $p_{22} = (.95 + .05m - .5ma)T$. In order for the game in table 4 to be a stag hunt, four conditions must be met:

[1] $p_{11} > p_{21}$ (successful stag hunting is better than unilateral hare hunting)

[2] $p_{21} \geq p_{22}$ (unilateral hare hunting is at least as good as both hunting hare)

[3] $p_{21} > p_{12}$ (unilateral hare hunting is better than unilaterally hunting stag)

[4] $p_{21} + p_{22} > p_{11} + p_{12}$ (hare hunting is risk dominant)

Conditions [2] and [3] are always met, because m, a, and T are positive. Condition [1] is met when $T > \frac{10a}{10a-1}$ and condition [4] is met when $T < \frac{10a+1}{10a-1}$. To summarize then, assuming honest communication and no punishment, repetition transforms the Sir Philip Sidney game into a stag hunt when:

$$\frac{10a}{10a-1} < T < \frac{10a+1}{10a-1} \tag{1}$$

How many expected rounds of repetition are needed depends on the benefit the resource confers on a person in need.

Figure 2 presents a summary of a helpful categorizations of this game for different values of T and a, which will give a sense of which parameter values give rise to a stag hunt and which will be helpful in discussing the results below. The black band represents the case where equation 1 holds, and the game can

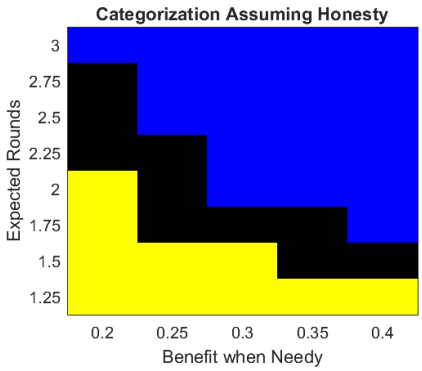


Figure 2: Categorization of the game, assuming honest communication and no punishment. Black indicates a classic stag hunt, where hunting hare is risk dominant. Blue indicates a stag hunt where hunting hare is not risk dominant. Yellow indicates that hunting hare is (weakly) dominant.

be categorized as a stag hunt. When $T > \frac{10a+1}{10a-1}$, in the top-right area, hunting hare is no longer risk dominant, and so the game would no longer fit the classical definition of a stag hunt (though, we may think of it as a sort of assurance game similar to the stag hunt). When $\frac{10a}{10a-1} > T$, in the bottom-left area, hunting hare is either dominant or weakly dominant, and so the game is no longer a stag hunt. In fact, when hunting hare is dominant, the game is a prisoners' dilemma.

For the results presented below, we will look at how the evolution of cooperation depends on both T and a (in the same range as in figure 2) when we no longer assume honest communication from the start, and then when we add the possibility of punishment.

3.2 Cooperation, communication, and punishment

We will now look at the co-evolution of communication and donation, to see how the addition of punishment for lying, i.e. claiming to be needy when not actually in need, affects the evolution of both honest communication and cooperation. To model this situation, we consider a Sir Philip Sidney game and add the possibility of punishment. If a healthy individual signals they are needy, there is some chance, e, that the receiver will discover the lie. If a lie is discovered, there is an option for the receiver to pay a cost c to inflict a punishment l on the liar.[11]

There are many possible strategies in this game. In order to keep the analysis tractable, we will consider a subset of these that are particularly relevant to the question at hand. We will consider two possible sender strategies:

[H] Honest, which sends the signal only when needy

[11] The detection of lies in this model may be thought of as akin to the 'incongruence hypothesis' in biology, where a receiver can detect a mismatch between sender type and signal. See Tibbetts [2013] for an overview.

[L] Liar, which always sends the signal

and three possible receiver strategies:

[R] Reciprocator, which plays TFT

[P] Punisher, which plays TFT and punishes whenever they detect a lie

[D] Defector, which ignores any signal and never donates

Both R and P will decide not to donate when they detect a lie, but P will additionally pay c to inflict a punishment l.[12] This punishment strategy is weakly dominated; R always does at least as well and is sometimes better than P.[13]

Each player has both a sender and a receiver strategy, leading to six types in the population. Evolution takes place according to the discrete-time replicator dynamics [Weibull, 1997], also used in Vanderschraaf [2016], by which strategies which are doing better than average at time t increase in frequency at time $t+1$:

$$x_i(t+1) = \frac{F_i x_i(t)}{F} \tag{2}$$

F_i is the average payoff for agents using strategy i, F is the average payoff in the population, and $x_i(t)$ is the proportion of agents using strategy i in generation t. Simulations of the evolutionary process were run for 3,000 time periods, with random starting frequencies of the possible strategies for each run. For each combination of parameters, 500 runs of the simulations were conducted in order to estimate the likely evolutionary outcomes.

Results will be shown using heatmaps, which capture the proportion different strategies take up at the end of the evolutionary process, averaging over all runs of the simulation. This gives a good overall picture of evolutionary outcomes but leaves out some information about, for instance, whether 40% of simulations converged to entirely honest signaling or whether all simulations converged to roughly 40% honesty. So, these heatmaps will give a rough visual representation of outcomes, which will be supplemented in-text with details about whether populations converge to particular outcomes. Note that the heatmaps capturing the proportion of punishers range from 0 to .5, while the heatmaps for honesty and defection range from 0 to 1.

As in Vanderschraaf [2016, 2018], in order to investigate whether punishment increased both honesty and cooperation, we compare the cases where c and l are both set to 0 (that is, there is no cost to punish and no penalty, so punishment effectively does not exist) to cases where there is some positive c and l.

[12] If the receiver does not donate after detecting a lie, it is assumed that the sender does not interpret this as a failure to cooperate because they understand the situation is different from a receiver not donating to a person in need. Additionally, a lie is not considered the same as failure to cooperate; all that matters for future donation is whether an individual donated in previous interactions. Therefore, reciprocal donation can continue despite lies being detected.

[13] The number of strategies has been restricted in order to keep the analysis tractable. Future research may consider what happens when seemingly counter-intuitive strategies are included, e.g. when senders never signal even when they are needy. Seemingly counter-intutive strategies may be more evolutionarily important than one initially suspects (see, e,g, Skyrms [2014]).

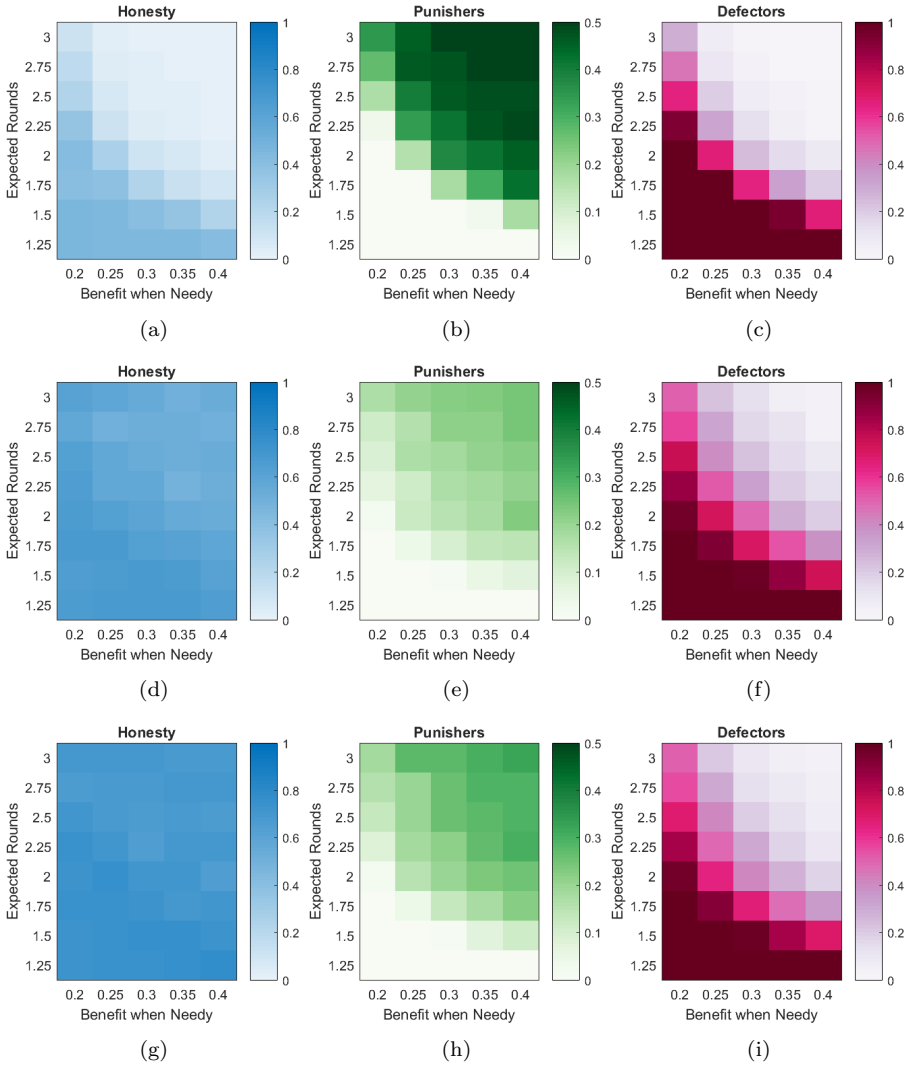


Figure 3: Evolution of sender and receiver strategies with (a-c) no punishment, (d-f) punishment where $c = .1$ and $l = .2$, and (g-i) punishment where $c = .1$ and $l = .3$. Results are given in terms of the proportion of the poppulation different strategies take up at the end of the evolutionary process, averaging over all runs of the simulation.

Figures 3a-3c show results for the no punishment case ($c = l = 0$), where $m = .3$ and $e = .9$. For this combination of parameters, we look at how T and a affect the evolution of various sender and receiver strategies. Figure 3a shows the average proportion of honest senders left at the end of simulations. Since there are only two possible sender strategies, the proportion of liars can be inferred. As figure 3a shows, without punishment, it is unlikely for honesty to evolve. With no punishment for lying, always sending the signal has no drawbacks and is advantageous against R and P senders who only donate when they get the signal. The greatest proportion of honest senders occurs for low values of T and a, which can be explained by looking at the evolution of receiver strategies. In figure 3c, we can see that there are also more defectors for these low T and a values. That is, receivers ignore the signal and do not donate any resources. Since H and L get the same payoff when communication is simply ignored, in these cases, evolution tends not to converge on one sender strategy or the other, but to lead to a mix of H and L.[14]

Looking more closely at the evolution of receiver strategies, we can compare figures 3b and 3c, which show the evolution of P and D, respectively (while the proportion of R can be inferred). In those cases where defection does not take over (top-right, high T and a values), cooperation takes over. In these cases, there is some combination of P and R, averaging around a 50/50 split. Of course, this makes sense because when the cost to punish and penalty for lying are both zero, R and P are, in effect, the same strategy. Whether we get cooperation (some combination of R and P) or defection is roughly predicted by the categorization of the game according to T and a as shown in figure 2. Defection evolves in those cases where hare hunting is dominant, i.e. when $\frac{10a}{10a-1} > T$. In the cases where the game can be categorized as a stag hunt, cooperation generally evolves less than half the time.[15] When stag hunting is risk dominant, i.e. $T > \frac{10a+1}{10a-1}$, cooperation is more likely to evolve.

What happens when we add punishment? Figures 3d-3f show results for $c = .1$ and $l = .2$. In figure 3d, we can see that the average proportion of honest senders left at the end of simulations is roughly 65% and does not depend on T or a, though looking at average proportion hides some trends. In particular, in the top-right area of the figure (high T and a), the population is more likely to converge to a population of all honest senders. For example, when $T = 3$ and $a = .4$, the population converges to all honesty roughly 50% of the time. This is compared to the bottom-left area (low T and a), where the population is more likely to not converge to one or the other strategy. This is consistent with the explanation of the results shown in figure 3a, where defection taking over as the

[14] Less than half the population is honest in these cases because there is selection against H before R and P die out.

[15] There are a few cases where cooperation evolves more than half the time, but this is consistent with the game being a stag hunt because cooperative strategies are over-represented at the start of simulations. At the start of each simulation, the initial distribution of strategies is chosen uniformly at random from the range of possible starting points to estimate likelihood of different evolutionary outcomes. However, since R and P are considered two different strategies, there are two possible cooperative strategies, and so on average will comprise 2/3 of the initial population.

receiver strategy allows a mix of sender strategies to persist.[16]

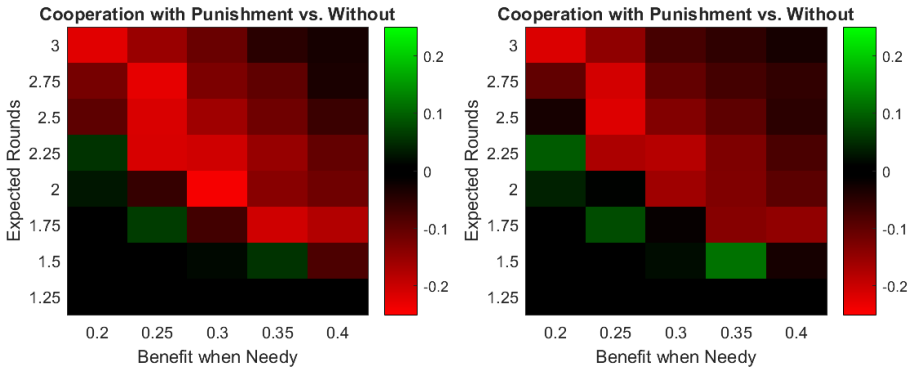
Figures 3e and 3f show the effect of adding punishment on the evolution of receiver strategies with $c = .1$ and $l = .2$. Unsurprisingly, comparing figures 3b and 3e shows that making punishment costly to perform decreases the average proportion of P left at the end of simulations. On the other hand, the effect on the evolution of D is more surprising: adding punishment actually increases the amount of defection, particularly in the middle region of the heatmap. This result will be discussed in more detail shortly.

Figures 3g-3i show results for when punishment is slightly more effective: we keep $c = .1$, but increase l to .3. In this case, we end up with more H on average, roughly 75%, as figure 3g shows. Again, the population tends to converge to an all honest population towards the upper-right (roughly 67% of the time for $T = 3$ and $a = .4$), while being more likely to end up with a mix of H and L as we move toward the bottom-left. Looking at figure 3h, we can see that there is more punishment when $l = .3$ at the end of simulations, especially for high T and a values. While punishment costs the same for figures 3e and 3h, we see more punishment in 3h because L is selected against more strongly due to suffering higher penalties for lying. That is, while punishers do not suffer less of a cost, they have to suffer it less often as the liars they punish disappear more quickly from the population.

The effects on the evolution of cooperation are more complicated in this case and harder to discern by just looking at figure 3. In some areas, there is more defection, in other areas there is less. To get a better picture of these effects, we can look at figure 4, which captures how the evolution of cooperation is affected by adding punishment of different forms. This figure summarizes how much more cooperation is expected when we add punishment by subtracting the proportion of cooperative (R and P) strategies in the no punishment case from the proportion of cooperative strategies in the punishment case. Positive numbers, depicted as green areas, indicate there is more cooperation when there is punishment, and negative numbers, depicted as red areas, indicate there is less cooperation when there is punishment. Figure 4a shows that adding punishment with $c = .1$ and $l = .2$ generally does not lead to more cooperation, but often leads to less. Figure 4b shows that adding punishment with $c = .1$ and $l = .3$ has a similar effect, though there are more cases where punishment can increase cooperation to a larger degree.

Why does adding punishment often make it harder for cooperation to evolve? One might expect that punishment would encourage honesty and that honesty would encourage cooperation – cooperators donate less often, and only when the donation does the greatest good (when the sender is needy and so gets a larger benefit from the donation). It might seem as though this should make cooperation as a whole more effective. Instead, what happens is that, even when the cost to punish is low, the costly punishing of every lie outweighs any

[16] Even in those cases where defection eventually takes over, evolution still sometimes converges to honesty (e.g., roughly 11% of the time for $T = 1.25$ and $a = .2$). Even though P eventually disappears in these cases, enough punishers may survive for a long enough time for L to die out.



(a) (b)

Figure 4: How punishment affects the amount of cooperation when (a) the penalty for deception is .2 and (b) the penalty for deception is .3

gain in efficiency honest communication might confer. For example, if R or P is matched with D, they only potentially pay the cost of their altruistic behavior on the first round of the interaction. By contrast, a punisher interacting with a cooperator who lies about when they are needy may end up paying the cost to punish multiple times, while the cooperator who lies correspondingly incurs the punishment multiple times. So the punishment of lying promotes honesty, but as it does so it creates roadblocks for cooperation by making it less effective, and we get more chances for defection to evolve.

There is the further question of why adding punishment of lying sometimes leads to more cooperation and sometimes does not. While this game is quite complicated, we can get a sense of what factors are important by looking back at the categorization in figure 2. The cases where punishment impedes cooperation are cases where reciprocal donation is risk dominant, meaning it is likely to evolve (assuming honest communication). Those cases where punishment promotes cooperation are those cases where it would have been more difficult for cooperation to evolve, even assuming honest communication, because either the game is a stag hunt or hare hunting is weakly dominant. In these cases, having people always claim to be in need would only serve to benefit the defectors even more (as they now always benefit from cooperators in the first round, not just when they are needy), making it even less likely for cooperation to evolve. So, honesty is important to ensuring cooperation has a chance to evolve compared to defection, and adding punishment of lying can ensure cooperation has that chance.

It is important to note that the conclusion that punishment of lying can create roadblocks to cooperation may be sensitive to the fact that we are looking at how costly punishment, in particular, affects the evolution of cooperation. For example, if punishment takes the form of withholding future donations (as described in section 3 above), defection may not gain this sort of advantage as both

defectors and cooperative liars would similarly lose out on future cooperation.[17] We might also consider punishment in the sense that is included in models of indirect reciprocity, where one's cooperation is reciprocated, not necessarily be the person they are currently interacting with, but through cooperation by a future interactive partner. Punishment would then be carried out by a third party refusing to cooperate with someone who has failed to cooperate in the past. Others have shown that when factors like reputation or standing within a group affect whether a person receives future donations, cooperation can evolve through indirect reciprocity [Nowak and Sigmund, 1998, 2005, Panchanathan and Boyd, 2003, Ohtsuki and Iwasa, 2006].

Of course, indirect reciprocity may be thought to include an aspect of communication as well, where people have to communicate about previous interactions in order to determine what type of person they are interacting with now. Models have shown that cooperation and communication about who cooperates – conceptualized either as 'rumors' [Nakamaru and Kawata, 2004] or 'moral signals' [Smead, 2010] – can co-evolve. Many of these models include similar complexities to the models presented here, where whether such punishment promotes cooperation depends on how punishment is implemented. For instance, errors in the perception of reputation can undermine indirect reciprocity, as can failing to prevent punishers (those who withhold refuse to cooperate as means to punish those who 'deserve' the punishment) from being punished themselves [Panchanathan and Boyd, 2003, Ohtsuki and Iwasa, 2006].

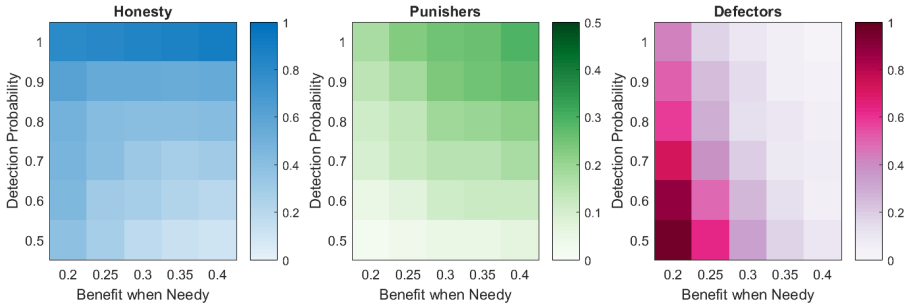
3.2.1 Lie detection probability

One worry one might have about the previous results is that $e = .9$, i.e. lie detection is very good. The results in figure 5 show what happens when we vary the benefit when needy and the lie detection probability, assuming $c = .1$ and $l = .2$, and $T = 3$. For context, this means that the top row of figures 3d-3f is now the second row of figures 5a-5c.

While better lie detection leads to more honesty overall, as figure 5a shows, honesty can still emerge as the lie detection probability decreases. Better lie detection matters more to the evolution of honesty when the benefit when needy is higher; low lie detection rates make less of a difference when the population ends up with mostly D, and, as figure 5c shows, D evolves more often when there is lower lie detection. As in figure 3, in the cases where deception evolves (lower left corner), the population is less likely to converge to one or another sender strategy, but rather end at a mix of H and L.

Punishment evolves more often as lie detection is better, as figure 5b shows. This might seem to conflict with the reasoning above, because better lie detection means the punishers must pay the cost of punishing more often, putting them at evolutionary disadvantage compared to R and D receiver strategies. However, since L suffers more than P from each punishment (i.e., $l > c$), more effective lie detection means faster elimination of L, and reduced time the pun-

[17] Thanks to Nick Shea for this point.



(a) (b) (c)

Figure 5: Evolution of sender and receiver strategies where $c = .1$ and $l = .2$, and $T = 3$.

ishers need to pay the cost. This is a similar, thought not identical, reason to why increasing l can allow more punishment. In this case P must pay an increased cost to inflict more punishment, but since liars are affected more than punishers, this does not affect the fitness of P too much for too long, ultimately allowing cooperation (R and P strategies) to evolve.

In other words, the important comparison here is not P vs. D, but P vs. L. In the no punishment case, L always takes over, but cooperation can still evolve. In cases where there's punishment, cooperation is more likely when P is not too negatively affected by the cost of punishment. In this case, increasing lie detection increases the frequency of punishments, meaning initially P pays more often, but in the longer term P pays less often as L disappears more quickly from the population.

3.3 Punishing defection

In order to get a full understanding of the contrast between these results and those of Vanderschraaf [2016], we ought to compare the effects of punishing lying with the effects of punishing defection in the repeated Sir Philip Sidney game. Does punishment of defection have similar effects in the repeated Sir Philip Sidney game as it did in the stag hunt, or is there something about the differences in the game structure that leads to the surprising results in section 3.2? We will consider results for $c = .1$ and $l = .2$, and compare to the results in section 3.2 with the same parameter values. Figure 6 shows what happens when we augment the game to this possibility of punishing defection.

Figures 6a-6c show the evolution of sender and receiver strategies when we allow punishment of defection, but not punishment of lying. As we can see in figure 6a, very little honesty is expected in this case, as honest senders can only persist when defection takes over as the receiver strategy, i.e. when selection on sender strategies disappears. Punishers can persist whenever cooperative strategies take over, when there is no one left to punish and no selection against

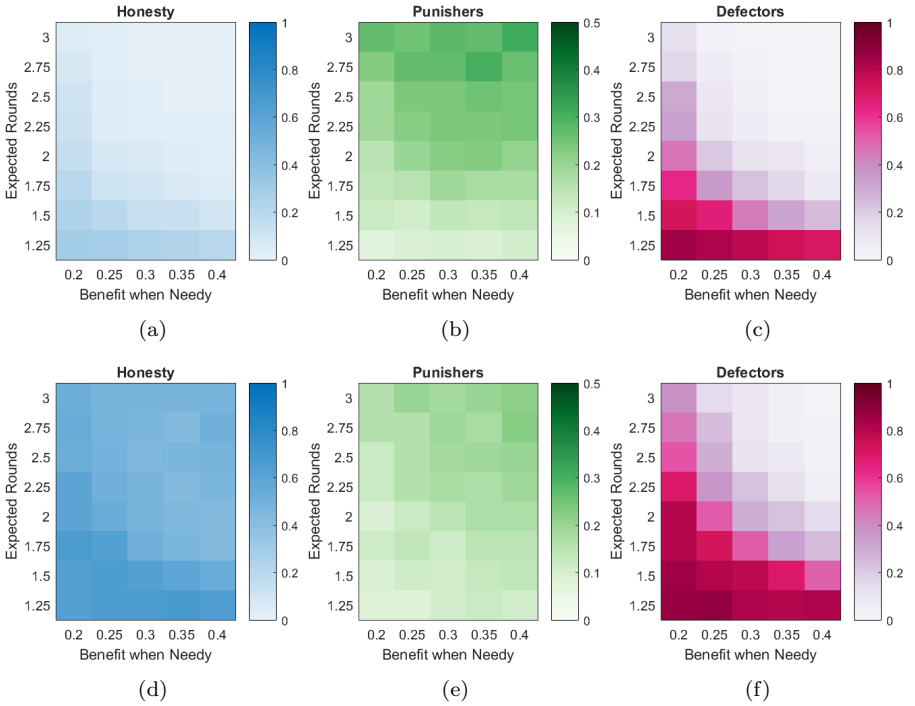
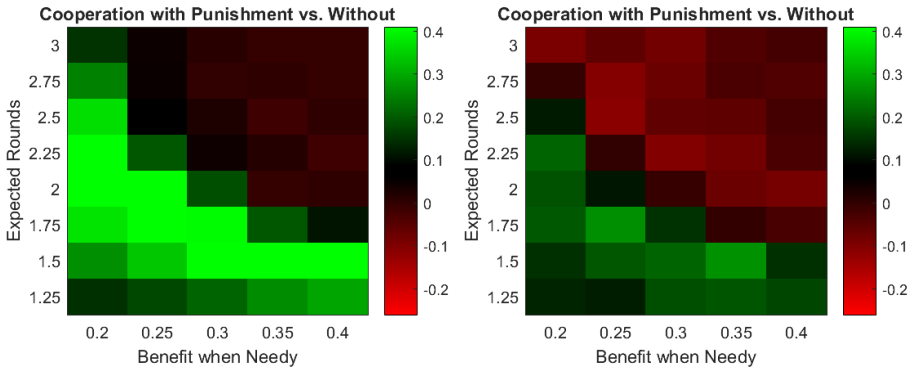


Figure 6: Evolution of sender and receiver strategies, where $c = .1$ and $l = .2$, with (a-c) punishment of defection only, and (d-f) punishment of both lying and defection

punishers. In comparison with figure 3c (the no punishment case) there is much less defection expected at the end of simulations. Figure 7a summarizes this difference between figures 6c and 3c. We can see that punishment of defection has much the same effect as it did in the stag hunt: it generally increases the basin of attraction for cooperation. (Note the change of scale in figure 7 compared to figure 4, as punishment of defection was able to promote cooperation to greater extent.[18])

We might also be interested in the effects of combining the two types of punishment: are the results intermediate between the effects of each type of punishment separately, or is there some sort of interactive effect, etc.? Figures 6d-6f show the evolution of sender and receiver strategies when we allow both punishment of lying and punishment of defection. Comparing figure 3d and figure 6d, we see similar amounts of honesty with punishment of lying only compared to punishment of both lying and defection. Comparing figure 3e and figure 6e, we can see that there are similar amounts of punishers in those

[18]In fact, some of the brightest areas exceed the highest value in the scale, e.g., at $T = 1.75$ and $a = .25$, there is roughly 65% more cooperation with punishment. However, further increasing the maximum value of the scale suppresses variations in other areas of the heatmap, making comparisons with other figures more difficult (and these comparisons are more important to the argument of the paper).



(a) (b)

Figure 7: How punishment affects the amount of cooperation with (a) punishment of defection only and (b) punishment of both lying and defection

cases where cooperation tends to evolve (toward the top-right). By contrast, there are more punishers in the bottom-left in figure 6e: these are the cases where defection nearly always evolves with punishment of lying only (figure 3f), while it evolves often, but not always, when there is additionally punishment of defection (figure 6f).

Looking at figure 7b, we can see that the evolution of cooperation when there is punishment of both lying and defection does seem to be intermediate between the punishment of lying only (figure 4a) and punishment of defection only (figure 7a) cases. There are more cases where punishment of both improves chances of cooperation, compared to punishment of lying only, but there are still a substantial number of cases where punishment hurts successful cooperation.

4 In a weakly dominated strategy there may or may not be strength

If we think of justice as arising out of the mutual advantage gained from certain kinds of conventions, this leads to an increased importance of understanding of the emergence and function of conventions regulating our behavior. What sorts of factors promote the evolution of beneficial social coordination, and which may prevent it? In particular, does punishment, as is commonly assumed, support social coordination, to the mutual benefit of all?

I have argued that when there are multiple stages in a cooperative endeavor, augmenting a game to include a weakly dominated punishment strategy may or many not promote social coordination. Punishment at one stage can make it harder for the cooperative endeavor as a whole to succeed. This runs counter to the received wisdom that punishment promotes cooperation, which has been found to be true in games that study only altruism, or only cooperation, without

including the communication aspect which often precedes those behaviors in real world conventions.

It is, of course, necessary to simplify real world situations in order to gain any understanding of the possible evolution of the conventions that regulate our behavior. We simply cannot include all the details and expect to grasp what the important factors are. This means we often investigate individual parts of a more complicated convention separately, in order to reveal basic insights. However, the conclusions we draw based on these parts, separately, may not hold when we start to think about how the parts fit together in a more complicated situation. This is similar to a point made by Wagner [2012], though he finds a somewhat more encouraging result that cooperation and fair division are more likely to emerge when we consider a compound stag hunt/demand game that combines both aspects of joint labor. In all, our understanding of what factors lead to mutually beneficial conventions may change when we begin to appreciate the complexity that is often involved in successful coordination.

References

J. Bednar and S. Page. Can game (s) theory explain culture? the emergence of cultural behavior within multiple games. *Rationality and Society*, 19(1): 65–97, 2007.

S. Bowles and H. Gintis. The evolution of strong reciprocity: cooperation in heterogeneous populations. *Theoretical population biology*, 65(1):17–28, 2004.

S. Bowles and H. Gintis. *A cooperative species*. Princeton University Press, 2011.

R. Boyd and S. Mathew. Third-party monitoring and sanctions aid the evolution of language. *Evolution and Human Behavior*, 36(6):475–479, 2015.

R. Boyd and P. J. Richerson. *Culture and the evolutionary process*. University of Chicago press, 1988.

R. Boyd and P. J. Richerson. Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and sociobiology*, 13(3):171–195, 1992.

R. Boyd, H. Gintis, S. Bowles, and P. J. Richerson. The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences*, 100(6):3531–3535, 2003.

J. P. Bruner and H. Rubin. Inclusive fitness and the problem of honest communication. *The British Journal for the Philosophy of Science*, 71(1):115–137, 2020.

D. Catteeuw, T. A. Han, and B. Manderick. Evolution of honest signaling by social punishment. In *Proceedings of the 2014 Annual Conference on Genetic and Evolutionary Computation*, pages 153–160, 2014.

A. K. Chavez and C. Bicchieri. Third-party sanctioning and compensation behavior: Findings from the ultimatum game. *Journal of Economic Psychology*, 39:268–277, 2013.

E. Fehr and U. Fischbacher. Third-party punishment and social norms. *Evolution and human behavior*, 25(2):63–87, 2004.

J. H. Fowler. Altruistic punishment and the origin of cooperation. *Proceedings of the National Academy of Sciences*, 102(19):7047–7049, 2005.

J. Gale, K. G. Binmore, and L. Samuelson. Learning to be imperfect: The ultimatum game. *Games and economic behavior*, 8(1):56–90, 1995.

H. Gintis. Strong reciprocity and human sociality. *Journal of theoretical biology*, 206(2):169–179, 2000.

J. Henrich and R. Boyd. Why people punish defectors: Weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *Journal of theoretical biology*, 208(1):79–89, 2001.

D. Hume. *An Enquiry Concerning the Principles of Morals: A Critical Edition*. Clarendon Press, 2000a.

D. Hume. *A Treatise of Human Nature*. Oxford University Press, 2000b.

D. Lewis. *Convention: A Philosophical Study*. Harvard University Press, 1969.

S. Mathew, R. Boyd, and M. Van Veelen. Human cooperation among kin and close associates may require enforcement of norms by third parties. *Cultural evolution*, pages 45–60, 2013.

J. Maynard Smith. Honest signaling, the philip sidney game. *Animal Behavior*, 42:1034–1035, 1991.

M. Nakamaru and M. Kawata. Evolution of rumours that discriminate lying defectors. *Evolutionary Ecology Research*, 6(2):261–283, 2004.

M. A. Nowak and K. Sigmund. Evolution of indirect reciprocity by image scoring. *Nature*, 393(6685):573–577, 1998.

M. A. Nowak and K. Sigmund. Evolution of indirect reciprocity. *Nature*, 437 (7063):1291–1298, 2005.

H. Ohtsuki and Y. Iwasa. The leading eight: social norms that can maintain cooperation by indirect reciprocity. *Journal of theoretical biology*, 239(4): 435–444, 2006.

K. Panchanathan and R. Boyd. A tale of two defectors: the importance of standing for evolution of indirect reciprocity. *Journal of theoretical biology*, 224(1):115–126, 2003.

P. Rich and K. J. Zollman. Honesty through repeated interactions. *Journal of theoretical biology*, 395:238–244, 2016.

P. J. Richerson and R. Boyd. *Not by genes alone: How culture transformed human evolution*. University of Chicago press, 2008.

L. Samuelson. Evolution and game theory. *Journal of Economic Perspectives*, 16(2):47–66, 2002.

B. Skyrms. *The stag hunt and the evolution of social structure*. Cambridge University Press, 2004.

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

B. Skyrms. *Evolution of the social contract*. Cambridge University Press, 2014.

R. Smead. Indirect reciprocity and the evolution of "moral signals". *Biology & philosophy*, 25(1):33–51, 2010.

E. A. Tibbetts. The function, development, and evolutionary stability of conventional signals of fighting ability. *Advances in the Study of Behavior*, 45: 49–80, 2013.

P. Vanderschraaf. In a weakly dominated strategy is strength: Evolution of optimality in stag hunt augmented with a punishment option. *Philosophy of Science*, 83(1):29–59, 2016.

P. Vanderschraaf. *Strategic justice: Convention and problems of balancing divergent interests*. Oxford University Press, 2018.

E. O. Wagner. Evolving to divide the fruits of cooperation. *Philosophy of Science*, 79(1):81–94, 2012.

J. W. Weibull. *Evolutionary game theory*. MIT press, 1997.

K. J. Zollman. Explaining fairness in complex environments. *politics, philosophy & economics*, 7(1):81–97, 2008.