Journal of Theoretical Biology 421 (2017) 189–202

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/jtbi

Third-party punishment as a costly signal of high continuation probabilities in repeated games

Jillian J. Jordan[a,*], David G. Rand[a,b,c]

[a] Department of Psychology, Yale University, 2 Hillhouse Avenue, New Haven, CT, 06511, United States
[b] Department of Economics, Yale University, United States
[c] School of Management, Yale University, United States

ARTICLE INFO

Article history:
Received 13 October 2016
Revised 6 March 2017
Accepted 4 April 2017
Available online 5 April 2017

Keywords:
Cooperation
Trust
Direct reciprocity
Reputation
Evolution
Game theory
Dispersal

ABSTRACT

Why do individuals pay costs to punish selfish behavior, even as third-party observers? A large body of research suggests that reputation plays an important role in motivating such third-party punishment (TPP). Here we focus on a recently proposed reputation-based account (Jordan et al., 2016) that invokes costly signaling. This account proposed that “trustworthy type” individuals (who are incentivized to cooperate with others) typically experience lower costs of TPP, and thus that TPP can function as a costly signal of trustworthiness. Specifically, it was argued that some but not all individuals face incentives to cooperate, making them high-quality and trustworthy interaction partners; and, because the same mechanisms that incentivize cooperation also create benefits for using TPP to deter selfish behavior, these individuals are likely to experience reduced costs of punishing selfishness. Here, we extend this conceptual framework by providing a concrete, “from-the-ground-up” model demonstrating how this process could work in the context of repeated interactions incentivizing both cooperation and punishment. We show how individual differences in the probability of future interaction can create types that vary in whether they find cooperation payoff-maximizing (and thus make high-quality partners), as well as in their net costs of TPP – because a higher continuation probability increases the likelihood of receiving rewards from the victim of the punished transgression (thus offsetting the cost of punishing). We also provide a simple model of dispersal that demonstrates how types that vary in their continuation probabilities can stably coexist, because the payoff from remaining in one’s local environment (i.e. not dispersing) decreases with the number of others who stay. Together, this model demonstrates, from the group up, how TPP can serve as a costly signal of trustworthiness arising from exposure to repeated interactions.

1. Introduction

A defining feature of human social behavior is our willingness to punish selfish and immoral behavior. A considerable body of previous work has explored the evolution of punishment enacted by directly affected victims (Boyd et al., 2003; Fowler, 2005; Garcia and Traulsen, 2012; Hauert et al., 2007; Hilbe and Traulsen, 2012; Nakamaru and Iwasa, 2006; Nowak et al., 2000; Rand et al., 2010; Rand and Nowak, 2011; Rand et al., 2013; Roberts, 2013; dos Santos et al., 2010; Tarnita, 2015). Yet empirical evidence shows that people are willing to punish even as third-party observers who have not been directly harmed or affected in any way (Fehr and Fischbacher, 2004; Hamlin et al., 2011; Henrich et al., 2006; Jordan et al., 2014; Mathew and Boyd, 2011; McAuliffe et al., 2015; Rockenbach and Milinski, 2006). Such third-party punishment (TPP) has the consequence of promoting cooperation, because punishment deters selfish behavior (Balafoutas et al., 2014; Boyd et al., 2010; 2003; Boyd and Richerson, 1992; Charness et al., 2008; Fehr and Fischbacher, 2004; Jordan et al., 2015a; 2015b; Mathew and Boyd, 2011). However, it also has costs: for example, punishing typically involves time and effort, and can result in retaliation from the target of punishment (Balafoutas et al., 2014; Boyd et al., 2010). Why should disinterested third parties punish, especially in cases where the costs to the punisher outweigh any benefits he or she will derive from deterring future selfishness?

One important explanation concerns reputation: people may punish in order to earn social benefits from others. Consistent with this possibility, evidence suggests that observability increases third-party punishment (Kurzban et al., 2007). This may reflect systems of indirect reciprocity in which there are social norms for punishment (Panchanathan and Boyd, 2004), such that punishers

[*] Corresponding author.
E-mail addresses: jillian.jordan@yale.edu (J.J. Jordan), david.rand@yale.edu (D.G. Rand).

http://dx.doi.org/10.1016/j.jtbi.2017.04.004
0022-5193/© 2017 Elsevier Ltd. All rights reserved.

are rewarded (Raihani and Bshary, 2015b) or non-punishers are punished (Boyd and Richerson, 1992). Punishment may also confer reputation benefits because it serves to signal that the punisher will be cooperative or trustworthy towards others in the future (Barclay, 2006; Raihani and Bshary, 2015a), which is consistent with the observation that third-party punishers are trusted more in economic games (Barclay, 2006; Horita, 2010; Nelissen, 2008).

Costly signaling is one possible way such signaling could function. The key premise of costly signaling theory (Spence, 1973; Zahavi, 1975) is that seemingly unrelated signals can be used to convey information about partner quality, when quality is difficult to observe directly. Specifically, this can occur when individuals who make high-quality partners find it less costly to send signals than individuals who make low-quality partners, such that only high-quality individuals find it worthwhile to pay to signal (in exchange for the benefit of being chosen as a partner). As a result, an individual's quality can be inferred from his or her signaling behavior. A large literature (Bliege Bird et al., 2001; Boone, 1998; Gintis et al., 2001; Roberts, 1998; Smith and Bliege Bird, 2000; 2005; Wright, 1999; Zahavi, 1977; 1995) suggests that cooperative behavior may serve as a costly signal of partner quality. Much of this literature has focused on cooperation as a signal of traits that reduce the costs of conferring benefits on others (e.g. strength, skill, access to resources). Cooperation as a signal of willingness to confer benefits has also been discussed (Gintis et al., 2001; Smith and Bliege Bird, 2005). More recently, it has also been proposed that *punishment* could serve as a costly signal of partner quality (Barclay, 2006; Gintis et al., 2001; Raihani and Bshary, 2015a; Smith and Bliege Bird, 2005).

Here, we focus on a particular framework for modeling third-party punishment as a costly signal of quality—specifically, of trustworthiness—from recent paper by Jordan et al. (2016) (hereafter JHBR). The model proposed by JHBR seeks to address the question of *why* individuals who make trustworthy (and thus high-quality) interaction partners might experience lower costs of engaging in TPP. To do so, it starts from the premise that there are many well-established mechanisms for the evolution of cooperation (e.g direct reciprocity, indirect reciprocity, institutional reward and punishment, etc.) (Boyd and Richerson, 1992; Jordan et al., 2015a; 2015b; Nowak, 2006), and that there are stable individual differences in exposure to these mechanisms—such that some but not all individuals have incentives to cooperate, creating variation in partner quality. The model then combines the insight of costly signaling with the insight of deterrence theory (i.e. that punishment, when targeted at selfish norm violators, promotes cooperative behavior) to explain the evolution of TPP as a costly signal of trustworthiness.

The key idea behind the model is that the same mechanisms that incentivize some individuals to be cooperative (and thus make them trustworthy, high-quality partners) also provide incentives to deter selfishness via TPP, which offsets their costs of punishment. For example, in the same way that direct reciprocity (Trivers, 1971) can incentivize cooperation when repeated interactions are possible (I cooperate with you today so that you reward me tomorrow), it can also incentivize punishment (I punish when somebody harms you today, deterring others from harming you in the future—again, so that you will reward me tomorrow). So, if an individual is particularly likely to experience repeated interactions, and thus faces particularly strong incentives to cooperate with others (making her a very trustworthy, high-quality partner), she *also* is likely to derive larger reciprocity-based benefits from punishing on the behalf of others (who can reward her later), making TPP less net costly for her. This same logic could also work for other incentive mechanisms, such as rewards from institutions that encourage both cooperation and TPP. As such, engaging in TPP can function as an honest signal of exposure to mechanisms incentivizing cooperation (and thus partner quality). This signaling process may help explain why people engage in TPP, even when, on their own, deterrence-based benefits are too small to compensate the costs of punishing.

To model this process, JHBR used a general framework that employs a standard costly signaling setup (Gintis et al., 2001; Spence, 1973). In JHBR's model, a Chooser decides whether to accept a Signaler based on inferring the Signaler's type from costly signals sent by the Signaler. Being accepted by the Chooser is always beneficial for the Signaler, but accepting the Signaler is only beneficial to the Chooser if the Signaler is a high-quality type (otherwise, accepting is costly to the Chooser). JHBR showed how TPP, if it is less costly for high-quality (i.e. "trustworthy") types, can act as a costly signal of trustworthiness.

This framework can apply to a wide range of specific situations which give rise to (i) the general partner choice payoff structure described above and (ii) the differential costliness of TPP. To emphasize this generality, JHBR presented the basic framework and described various possible implementations that would satisfy requirements (i) and (ii) (such as the aforementioned direct reciprocity and institutional reward examples) at an abstract conceptual level.

Here, we complement JHBR's model by laying out a concrete implementation of where types originate and why they have different TPP costs, based on repeated interactions and direct reciprocity. We provide a detailed "from-the-ground-up" model that begins with differences between individuals in their probabilities of being present for future interaction (i.e. differences in continuation probability). We show how this leads to types that vary in their trustworthiness, and thus quality, as well as in their net costs of third-party punishment (with trustworthy types experiencing larger deterrence-based benefits of TPP that offset their costs).

But why would there be stable individual differences in the probability of being present for future interaction? Individual differences that provide the basis for costly signaling could come from a number of outside forces that create stable variation (Gintis et al., 2001). For example, individuals might have different probabilities of future interaction because they have different mortality rates or probabilities of dispersing outside of their local environment. In the context of dispersal, for example, probabilities of leaving one's environment could be determined by evolutionary mechanisms unrelated to cooperation or punishment (e.g. a frequency-dependent process where the payoff of staying in the local environment decreases with the proportion of the population that stays (Roff, 1975), such that some individuals evolve a low dispersal probability while others evolve a high dispersal probability). Alternatively, dispersal or mortality rates could be determined by the environment, and thus not be heritable (e.g. some individuals mature when local resources are scarce and decide to disperse; or some individuals get an illness that increases their probability of death)—such that they do not evolve at all. Any of these processes can explain the stable coexistence of types with different continuation probabilities that forms the input to our model; thus, for most of this paper, we assume fixed types without modeling the process giving rise to them. However, in Section 4, we provide a simple model of dispersal and show how it can output stable variation in type.

The rest of this paper thus proceeds as follows. In Section 2, we describe the payoff structure of the partner choice interaction and the signaling interaction in our concrete, direct-reciprocity-based model; and show how the payoffs from this specific model map onto the general payoff structure of the model from JHBR. In Section 3, we present results from equilibrium calculations and evolutionary simulations for a signaling equilibrium that occurs in this concrete model. In Section 4, we describe our dispersal model that gives rise to stably coexisting types. In Section 5, we explain how our model can be extended to include multiple imperfect signals. In Section 6, we conclude. Finally, in Appendix A, we describe an additional signaling equilibrium that can occur in our concrete model.

2. Model

2.1. Overview

Like in JHBR, we model a game in which pairs of agents interact, in the roles of "Signalers" and "Choosers". In each two-stage interaction, one Signaler is paired with one Chooser. First, in the signaling stage, Signalers have the opportunity to pay costs to send signals, which are observed by Choosers. Second, in the partner choice stage, Choosers decide whether to accept Signalers as partners on the basis of these signals. Here, we describe the interactions that take place in these two stages in more detail. We start by describing the partner choice stage (even though it occurs second), because it helps lay out the general framework (based on direct reciprocity) that underlies this concrete model. Then, we describe the signaling stage, which also involves play with a third kind of agent, a "Victim".

2.2. Partner choice stage

In the partner choice stage, Choosers and Signalers play an asynchronous repeated PD, where cooperating involves paying cost c to deliver benefit b, and defecting involves paying no costs to deliver no benefits. The Chooser makes the first move. If the Chooser defects in this opening move, reflecting the choice not to invest in a partnership with the Signaler (i.e. "rejects" the Signaler, in the terminology of the JHBR model), the game does not continue and there are no additional rounds. However, if the Chooser cooperates in the opening move, reflecting the choice to invest in a partnership with the Signaler (i.e. "accepts" the Signaler, in the terminology of the JHBR model), the game continues and there may be additional rounds.

Specifically, subsequent interaction between the Chooser and Signaler occurs only if both the Chooser and Signaler are present in the subsequent period. Each player has their own individual and fixed probability w of being available for future interaction with their current partner; thus, the probability that another round occurs between Chooser and Signaler is $w_c w_s$, where w_c and w_s are the continuation probabilities of the Chooser and the Signaler respectively. Thus, if the Chooser cooperates in the opening move, another round in which the Signaler moves occurs with probability $w_c w_s$; then another round in which the Chooser moves occurs with probability $(w_c w_s)^2$, and so on.

For simplicity, we assume that all players (be they Choosers or Signalers) either have a "high" or "low" probability of being present for future interaction; thus, w_c and w_s each can take one of two values: w_H and w_L, where $0 < w_L < w_H < 1$. We refer to individuals with the high probability w_H as "high types", and individuals with the low probability w_L as "low types". Furthermore, let h be the fraction of the population that is high type, and $1 - h$ be the fraction that is low type.

We now consider the expected payoffs of a Chooser and Signaler in the asynchronous PD partner choice stage. We assume that Choosers and Signalers each have two strategies available to them. First, Choosers decide between playing Reject (do not cooperate on the opening move and do not initiate an interaction with the Signaler) or Tit-for-tat (cooperate on the opening move, and then copy the Signaler's previous move for the duration of the game). Then, if the Chooser plays TFT, Signalers decide between also playing TFT (copy the Chooser's previous move for the duration of the game) or ALLD (always defect). We restrict the strategy set to these two strategies for tractability; see Conclusion for a discussion of how a fuller strategy space might affect our results.

If the Chooser (who moves first) plays TFT, and the Signaler responds by also playing TFT, the Chooser pays c in the first round (because the round occurs with certainty, and cooperating costs c), then receives $bw_c w_s$ in the second round (because the second round occurs with probability $w_c w_s$, and cooperation gives benefit b), then pays $c(w_c w_s)^2$ in the third round, then earns $b(w_c w_s)^3$ in the fourth round, and so on. This sum converges to $\frac{bw_c w_s - c}{1-(w_c w_s)^2}$. In contrast, the Signaler earns b in the first round, then pays $cw_c w_s$ in the second round, and so on, which converges to $\frac{b - cw_c w_s}{1-(w_c w_s)^2}$.

If the Chooser plays TFT, and the Signaler responds by playing ALLD, no payoffs are earned after the first round, and the Chooser ends up with $-c$, while the Signaler ends up with b.

Finally, if the Chooser plays Reject, the game does not continue (regardless of the Signaler's strategy), and both players earn nothing.

Thus we have the following payoff matrix for the partner choice stage:

	Signaler strategy	
Chooser strategy	TFT	ALLD
TFT	$\frac{bw_cw_s-c}{1-(w_cw_s)^2}, \frac{b-cw_cw_s}{1-(w_cw_s)^2}$	$-c, b$
Reject	0,0	0,0

(1)

where $0 < c < b$, and $0 < w_c, w_s < 1$.

From this payoff matrix, we see that it is always an equilibrium for the Chooser to play Reject and the Signaler to either play TFT or ALLD. Furthermore, it is an equilibrium for both players to play TFT when $w_c w_s > \frac{c}{b}$.

Using this setup for partner choice, we now turn to individual differences. Recall that low-type Choosers and Signalers have w_L probability of future interaction, whereas high types have w_H. When $\frac{c}{b} < {w_L}^2$, TFT is an equilibrium for all Signaler-Chooser pairs, and no signaling will occur. And when ${w_H}^2 < \frac{c}{b}$, TFT is never any equilibrium for any Signaler-Chooser pairs, and signaling will also never occur. In Appendix A, we consider the case in which ${w_L}^2 < \frac{c}{b} < w_L w_H$, such that TFT is an equilibrium when at least one player is a high type. In this case, high-type Choosers are willing to accept any Signalers, and low-type Choosers wish only to accept high-type Signalers. We demonstrate the existence of a signaling equilibrium under these conditions, in which high-type Signalers use TPP to signal their types to low-type Choosers, and high-type Choosers accept all Signalers.

Here, in the main text, we focus on the final case in which $w_L w_H < \frac{c}{b} < {w_H}^2$, such that TFT is only an equilibrium when both players are high types, and thus low-type Choosers are never willing to accept any Signalers, and high-type Choosers wish only to accept high-type Signalers. We demonstrate the existence of a signaling equilibrium in which high-type Signalers use TPP to signal their types to high-type Choosers, and low-type Choosers reject all Signalers.

To do so, we begin by evaluating Chooser play. If the Chooser is a high type, he should play TFT if and only if he believes the Signaler to also be a high type, and thus believes TFT to be an equilibrium. The reason is that the *Chooser* is likely to be present in the future to receive reciprocal cooperation from the Signaler, and so it is payoff-maximizing to invest in cooperating if and only if the *Signaler* is also likely to be present in the future to provide this reciprocal cooperation.

Conversely, if the Chooser is a low type, he should play Reject regardless of whether the Signaler is a high type or a low type. If he is a low type, TFT is never an equilibrium: he is unlikely to be present in the future to receive reciprocal cooperation from the Signaler, and so it is never worth the investment of opening with cooperation. Therefore, all interactions involving low-type Choosers produce zero payoff for both parties (and low-type Choosers can be ignored).

Now we turn to Signaler play. In the event that the Chooser plays TFT, the Signaler can infer that the Chooser is a high type (because TFT is never worthwhile for low types). Thus, she can expect the Chooser to be around in the future with high probability, and should respond with TFT if and only if she also is a high type, and is herself likely to be around in the future, in which case TFT is an equilibrium. When she is a high type, it is worth *her* investment in reciprocation for the chance of receiving future reciprocation; when she is a low type, it is better to exploit the Chooser's cooperation by not reciprocating, because she is unlikely to be around to receive further reciprocation.

We now consider the expected payoffs of the interaction of different types of Choosers and Signalers. Because it is never worth it for low-type Choosers to play TFT, all interactions involving low-type Choosers result in zero payoff for both players, and thus are irrelevant. The interaction of high-type Choosers (who may either play TFT or Reject) and Signalers of either high type (who will respond to TFT with TFT) or low type (who will respond to TFT with ALLD) gives the following payoff matrix:

	Signaler type	
Chooser Action	High	Low
TFT	$\frac{b{w_H}^2-c}{1-{w_H}^4}, \frac{b-{w_H}^2c}{1-{w_H}^4}$	$-c, b$
Reject	0,0	0,0

(2)

This game provides a specific implementation of the general partner choice payoff matrix from the main model of JHBR:

	Signaler type	
Chooser Action	Trustworthy	Exploitative
Accept	m, r_t	$-e, r_e$
Reject	0,0	0,0

(3)

where $0 < m, e, r$.

Thus, in the detailed model presented here,

$$\begin{aligned} m &= \frac{b{w_H}^2 - c}{1 - {w_H}^4} \\ e &= c \\ r_t &= \frac{b - {w_H}^2 c}{1 - {w_H}^4} \\ r_e &= b \end{aligned} \tag{4}$$

Note that in their model, JHBR set r_t (the payoff a trustworthy Signaler receives for being accepted) equal to r_e (the payoff an exploitative Signaler receives for being accepted), for simplicity. However, they noted in their SI that this need not be the case. In their model, there are two possible costs of punishing: a small cost s and a large cost ℓ. Their results hold so long as $s < r_t$, $r_e < \ell$, such that neither type of Signaler ever finds it worthwhile to pay the large signaling cost to be accepted by the Chooser, and both types of Signaler always find it worthwhile to pay the small signaling cost to be accepted by the Chooser.

2.3. Signaling stage

While Choosers cannot directly observe whether a Signaler is a high or low type, Signalers can send signals to provide information about their type. In this concrete model, we illustrate how TPP (in which a punisher punishes a transgressor who has acted selfishly toward a victim) can serve as a costly signal of type. For simplicity, we focus on the case where TPP costs are perfectly correlated with type, and TPP is the only available signal. Specifically, we derive the conditions under which it is an equilibrium for high-type (trustworthy) Signalers to engage in TPP and low-type (exploitative) Signalers to not engage in TPP, and for high-type Choosers to only accept Signalers who punished (i.e. conditions for the "Punishment Signaling" equilibrium to exist). In Section 5, however, we discuss how this model could be expanded to include multiple signals which are imperfect, such that some signals convey more information than others.

The model we present here is based on the premise that TPP deters the punished transgressor from acting selfishly towards the victim in the future. Therefore, TPP creates benefits for the victim, which the victim can then reciprocate by providing benefits to the punisher on a later occasion.

To capture this possibility of reciprocation, we model the Signaling stage as involving a modified asynchronous repeated PD played between the Signaler and the Victim. This modified PD, which we will call the "punishment PD", begins with a "punishment phase" in which the Signaler can punish on behalf of the Victim. Then, if the Signaler chooses to punish, the game moves to a "cooperation phase" for all future rounds. The cooperation phase is a standard asynchronous PD (as in the standard PD described above in the partner choice stage).

Thus, the Signaler is the first mover, and in the punishment phase chooses whether to punish on behalf of the Victim. Punishing causes the Signaler to incur a cost $c + k$ (which includes both the direct cost (e.g. resources, time, effort) of enacting punishment, and the expected cost arising from potential retaliation by the transgressor). Thus, k is the additional cost of punishing, over and above the cost of cooperating. Punishing delivers an expected benefit $b + j$ to the Victim (in the form of reduced future probability of the transgressor acting selfishly towards the victim). Thus, j is the additional benefit of punishing, over and above the benefit of cooperating.

If the Signaler decides not to punish, there is no interaction of any kind between the Signaler and the Victim, and both players earn zero in the Signaling stage of the game. However, if the Signaler decides to punish, the punishment PD then moves to a standard asynchronous repeated PD (the cooperation phase). In the first round of the cooperation phase, the Victim can reciprocate the Signaler's punishment by cooperating, which involves paying a cost c to deliver a benefit b to the Signaler (i.e. the same payoffs as in the partner choice PD). Then the Signaler has the opportunity to cooperate in this same way with Victim, then the Victim moves again, and so on. This cooperation phase is thus identical to the partner choice PD, with the Victim as the first mover; thus, in the cooperation phase, both the Victim and the Signaler can choose between playing TFT or ALLD.

In the punishment PD, the Signaler's punishment decision occurs with probability 1. Then, if the Signaler punishes, each future interaction occurs with probability $w_s w_v$, where w_s is the Signaler's probability of being present in the next round (the same probability as in the partner choice PD), and w_v is the Victim's. The values of w_s and w_v can again be w_H or w_L, based on the Signaler and Victim's types.

This gives the following payoff matrix for the punishment PD:

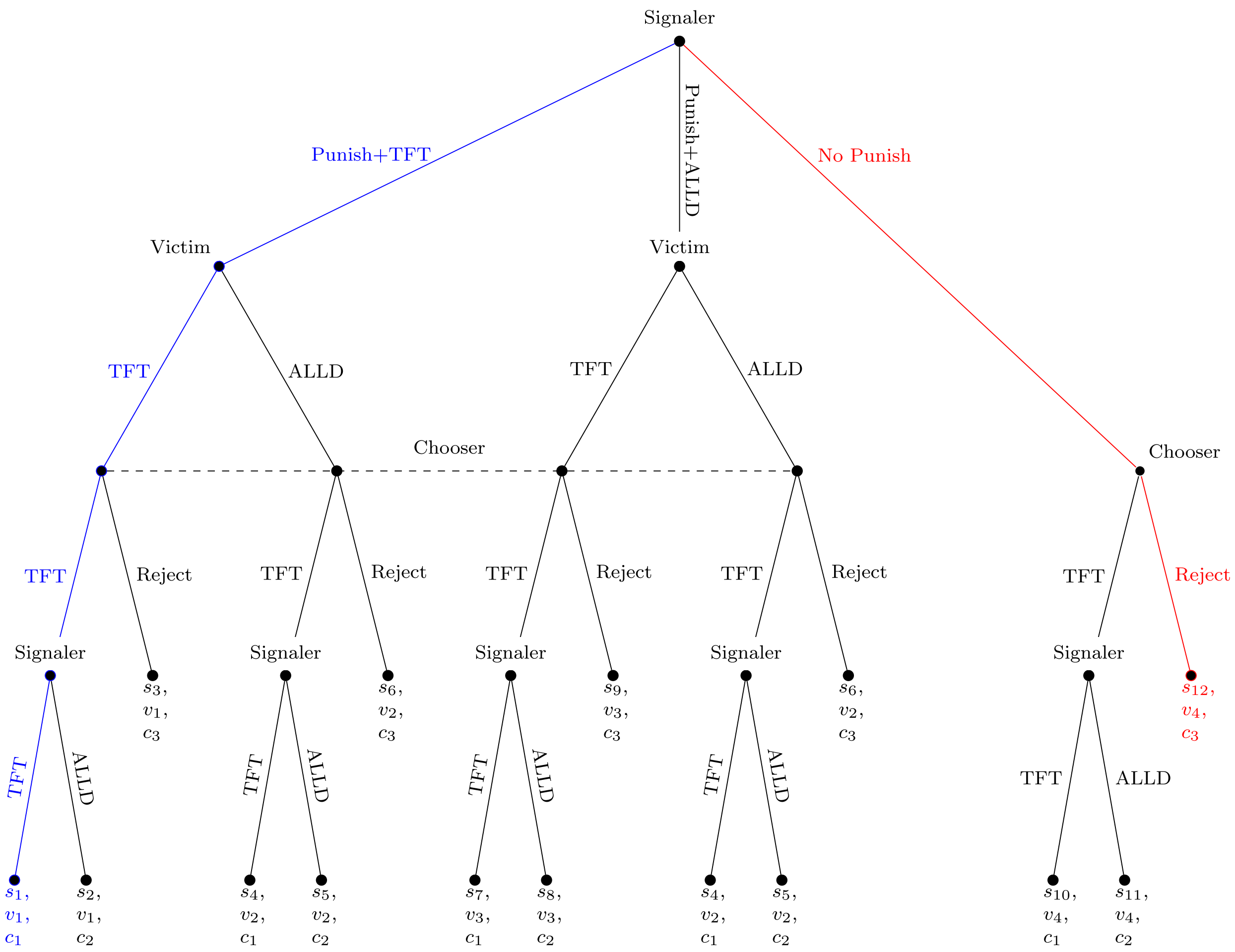
	Victim strategy	
Signaler strategy	TFT	ALLD
Punish/TFT	$\frac{bw_s w_v - c}{1-(w_s w_v)^2} - k, \frac{b - cw_s w_v}{1-(w_s w_v)^2} + j$	$-(c+k), b+j$
Punish/ALLD	$-(c+k) + bw_s w_v, b + j - cw_s w_v$	$-(c+k), b+j$
No Punish	0,0	0,0

(5)

Note that because we do not allow the Signaler to condition on the Victim's type, in expectation $w_v = hw_H + (1-h)w_L$.

3. Results

We now demonstrate the conditions for a Punishment Signaling strategy profile to be an equilibrium. This strategy profile is defined as follows. In the Signaling Stage (punishment PD), (A) high-



$$s_1 = \frac{bw_s w_v - c}{1-(w_s w_v)^2} - k + \frac{b - cw_c w_s}{1-(w_c w_s)^2} \qquad v_1 = \frac{b - cw_s w_v}{1-(w_s w_v)^2} + j \qquad c_1 = \frac{bw_c w_s - c}{1-(w_c w_s)^2}$$

$$s_2 = \frac{bw_s w_v - c}{1-(w_s w_v)^2} - k + b \qquad v_2 = b + j \qquad c_2 = -c$$

$$s_3 = \frac{bw_s w_v - c}{1-(w_s w_v)^2} - k \qquad v_3 = b + j - cw_s w_v \qquad c_3 = 0$$

$$s_4 = -(c+k) + \frac{b - cw_c w_s}{1-(w_c w_s)^2} \qquad v_4 = 0$$

$$s_5 = -(c+k) + b$$

$$s_6 = -(c+k)$$

$$s_7 = -(c+k) + bw_s w_v + \frac{b - cw_c w_s}{1-(w_c w_s)^2}$$

$$s_8 = -(c+k) + bw_s w_v + b$$

$$s_9 = -(c+k) + bw_s w_v$$

$$s_{10} = \frac{b - cw_c w_s}{1-(w_c w_s)^2}$$

$$s_{11} = b$$

$$s_{12} = 0$$

Fig. 1. Game in extensive form, with equilibrium path for the Punishment Signaling Equilibrium illustrated. In red, we show the equilibrium path when the Signaler is a low type, and the Victim and Choosers are of any type. In blue, we show the equilibrium path when the Signaler, Victim, and the Chooser are all high types. (For conciseness, we do not illustrate the equilibrium paths when the Signaler is a high type, but the Victim and/or Chooser is a low type.) (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

type Signalers play Punish/TFT; and (B) low-type Signalers play No Punish. In the Partner Choice stage (partner choice PD), (C) high-type Choosers play TFT if and only if the Signaler punished; (D) high-type Signalers play TFT; (E) low-type Choosers always play Reject; and (F) low-type Signalers always play ALLD. Fig. 1 illustrates the game in extensive form, as well as this equilibrium path.

We now derive the conditions required for there to be no incentive to deviate from each of these six features of the strategy profile, under the assumption discussed in Section 2.2 that

$$w_L w_H < \frac{c}{b} < {w_H}^2. \tag{6}$$

3.1. Signaling stage

3.1.1. (A) High-type Signalers play Punish/TFT

Starting with the signaling stage (i.e. the punishment PD), we begin by deriving the conditions required for (A) - that is, for high-type Signalers to not improve their payoff by deviating from playing Punish/TFT (given that on the equilibrium path, high-type Signalers must punish in order to elicit TFT play from high-type Choosers in the partner choice PD).

First, we calculate the total payoff to high-type Signalers in equilibrium. Calculating their payoff in the Signaling Stage requires specifying the Victim's punishment PD strategy. On the equilibrium path, if the Signaler opens the punishment PD by punishing in the punishment phase, the Victim can infer that the Signaler is a high type and therefore that the Signaler will play TFT in the cooperation phase (because on equilibrium path, low-type Signalers do not punish). From Eq. (5), observe that it is payoff-maximizing for the Victim to respond to a Signaler who plays Punish/TFT with TFT when $\frac{c}{b} < w_v w_s$. This condition is only satisfied when the Signaler and Victim both are high types (because $w_L w_H < \frac{c}{b} < {w_H}^2$). Thus, the Victim will respond to Punish/TFT with TFT if he is a high type (probability h), and will respond with ALLD if he is a low type (probability $1 - h$).

A high-type Signaler's expected payoff in the Signaling Stage from playing Punish/TFT is therefore

$$-k + h\left(\frac{b{w_H}^2 - c}{1 - {w_H}^4}\right) - (1 - h)c. \tag{7}$$

Then, in the Partner Choice Stage, high-type Signalers will receive $\frac{b - {w_H}^2 c}{1 - {w_H}^4}$ with probability h (when the Signaler is paired with a high-type Chooser, given that on the equilibrium path the high-type Chooser will play TFT because the Signaler punished, and the Signaler will respond with TFT); and 0 with probability $1 - h$ (when the Signaler is paired with a low-type Chooser, given that on the equilibrium path the low-type Chooser will play Reject). Thus, on the equilibrium path, across both stages a high-type Signaler playing Punish/TFT earns a total payoff of

$$-k + h\left(\frac{(b - c)(1 + {w_H}^2)}{1 - {w_H}^4}\right) - (1 - h)c. \tag{8}$$

For this behavior to be in equilibrium, the high-type Signaler must not be able to improve her payoff by switching to either Punish/ALLD or No Punish.

If the high-type Signaler deviates by playing Punish/ALLD, she earns an identical payoff in the Partner Choice Stage (because the Chooser's behavior in Partner Choice Stage is conditional only on whether the Signaler punishes). However, her expected payoff in the Signaling Stage becomes $-k + h(b{w_H}^2 - c) - (1 - h)c$, which is always lower than the Signaling Stage payoff of playing Punish/TFT (because $\frac{c}{b} < {w_H}^2$).

If the high-type Signaler deviates by playing No Punish, she earns zero payoff in both the Signaling Stage and the Partner Choice Stage (because on the equilibrium path, all Choosers respond to non-punishing Signalers with Reject in the partner choice PD). Thus a high-type Signaler cannot improve her payoff by deviating from Punish/TFT when

$$k < h\left(\frac{(b - c)(1 + {w_H}^2)}{1 - {w_H}^4}\right) - (1 - h)c. \tag{9}$$

3.1.2. (B) Low-type Signalers play No Punish

On equilibrium path, low-type Signalers play No Punish in the Signaling Stage, and are consequently always rejected in the Partner Choice Stage, earning a total payoff of 0. If the low-type Signaler deviates by instead playing Punish/ALLD, her expected payoff in the Signaling Stage becomes $-(c + k) + hbw_H w_L$, and in the Partner Choice Stage, the Signaler will earn b with probability h (when the Signaler is paired with a high-type Chooser, given that on the equilibrium path the high-type Chooser will play TFT because the Signaler punished, in response to which it is payoff maximizing for the low-type Signaler to play ALLD), and 0 with probability $1 - h$ (when the Signaler is paired with a low-type Chooser who will play Reject).

Thus, the low-type Signaler does not have an incentive to switch from No Punish to Punish/ALLD when

$$k > -c + hb(1 + w_H w_L). \tag{10}$$

This condition is sufficient to characterize when low-type Signalers have no incentive to deviate from No Punish, because Punish/TFT always earns a lower payoff for low-type Signalers than Punish/ALLD given that $w_L w_H < \frac{c}{b}$.

3.2. Partner choice stage

3.2.1. (C) High-type Choosers play TFT if and only if the Signaler punished

We now move to the partner choice stage. On the equilibrium path, a Signaler is a high type if and only if she punished in the Signaling Stage. From Eq. (1), we see that because $w_H w_L < \frac{c}{b} < {w_H}^2$, high-type Choosers playing against high-type Signalers (who punish) earn a higher payoff from sticking with TFT than from switching to Reject; and high-type Choosers playing against low-type Signalers (who do not punish) earn a higher payoff from sticking with Reject than from switching to TFT. Thus, high-type Choosers do not benefit from deviating from playing TFT if and only if the Signaler punished.

3.2.2. (D) High-type Signalers play TFT

On the equilibrium path, if the Chooser played TFT, he is a high type. From Eq. (1), we see that because $w_H w_L < \frac{c}{b} < {w_H}^2$, when both players are high types the Signaler earns a higher payoff from sticking with TFT than switching to ALLD. Thus, high-type Signalers do not benefit from deviating from playing TFT, as they only get to move if the Chooser also played TFT.

3.2.3. (E) Low-type Choosers play Reject and (F) Low-type Signalers play ALLD

From Eq. (1), we also see that because $w_H w_L < \frac{c}{b} < {w_H}^2$, if either player is a low type, neither Choosers nor Signalers can improve their payoffs by deviating from Reject and ALLD, respectively. Thus low-type Choosers and low-type Signalers have no incentive to deviate from always playing Reject and ALLD, regardless of the other player's behavior and inferred type.

3.3. Conditions for Punishment Signaling equilibrium

Taken together, the results of Sections 3.1 and 3.2 show that the Punishment Signaling strategy profile, where only high-type Signalers punish and high-type Choosers only accept Signalers who punish, is an equilibrium when

$$hb(1 + w_H w_L) < c + k < h\left(\frac{(b - c)(1 + {w_H}^2)}{1 - {w_H}^4} + c\right) \tag{11}$$

is satisfied, along with the assumption that $w_H w_L < \frac{c}{b} < {w_H}^2$ (such that cooperating in both the partner choice PD and the cooperation phase of the punishment PD is only payoff-maximizing when both players are high types). The condition given by Eq. (11) is the heart of the costly signaling model. It specifies that the cost of punishing $(c + k)$ is (i) sufficiently large that TPP is not worth it for low-type Signalers (whose TPP costs would not be offset by repeated cooperation with the Victim, because they are unlikely to be present in the future to receive such cooperation), even though TPP would elicit TFT from high-type Choosers in the Partner Choice Stage; but (ii) sufficiently small that TPP is worth it for high-type Signalers (whose TPP costs are offset by repeated cooperation in the Victim, because they are likely to be present in the future to receive such cooperation), in order to elicit TFT from high-type Choosers in the Partner Choice Stage. This condition can be simplified using the mapping to the main model parameters defined in Eq. (4), yielding

$$hb(1 + w_H w_L) < c + k < h(m + r_t + c). \tag{12}$$

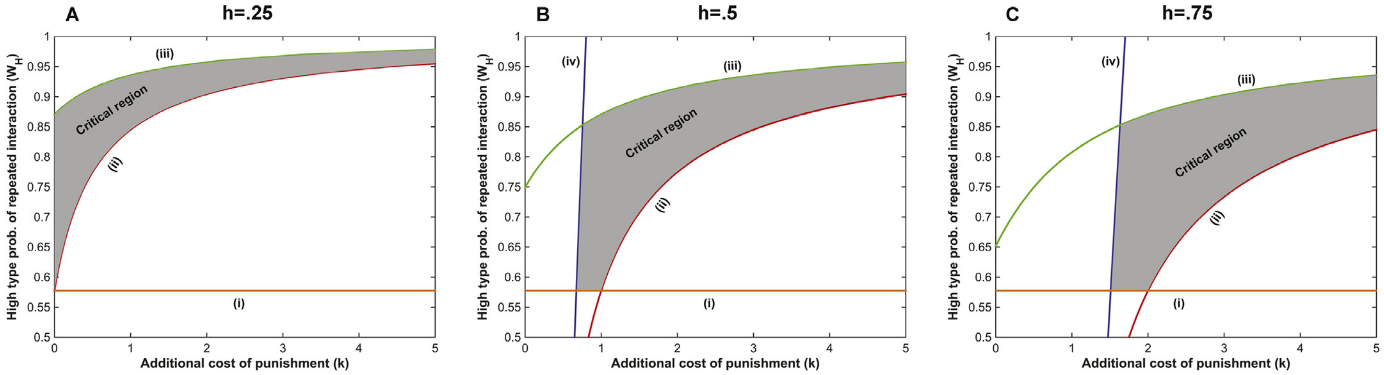


Fig. 2. Example parameter regions in which Punishment Signaling is an equilibrium, but no TPP will occur without signaling. In all panels, we set $w_L = .2$, $b = 3$, and $c = 1$. We then vary k on the x-axis and w_H on the y-axis, and h across panels. Line (i) in orange illustrates the minimum value of w_H for TFT to be an equilibrium in the Partner Choice stage when both players are high types. Line (ii) in red illustrates, for each value of k, the minimum value of w_H for Punish/TFT to be an equilibrium for high-type Signalers in the Signaling Stage, in a game with a Partner Choice Stage (i.e. for TPP to be worthwhile for such Signalers, given its signaling benefits). Line (iii) in green illustrates, at each value of k, the maximum value of w_H for No Punish to be an equilibrium for high-type Signalers in the Signaling Stage, in a game with no Partner Choice Stage (i.e. for no TPP to occur without Signaling). Line (iv) in blue illustrates, at each value of k, the maximum value of w_H for No Punish to be an equilibrium for low-type Signalers in the Signaling Stage, in a game with a Partner Choice Stage (i.e. for TPP to not be worthwhile for such Signalers, even given its signaling benefits). Thus the grey "critical region" indicates, for the relevant value of h, the set of values of k and w_H for which the Punishment Signaling strategy profile is an equilibrium, but no TPP will occur without signaling. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.4. Conditions for no punishment in the absence of signaling

A key claim of JHBR is that costly signaling can explain the occurence of third-party punishment in situations where it would *not* otherwise occur. Thus we are interested not only in the conditions required for (only) high-type Signalers to punish in the presence of signaling benefits (i.e. the Punishment Signaling equilibrium condition given by Eq. (11)), but also the conditions required for high-type Signalers to *not* punish in the absence of signaling benefits (i.e. when no Partner Choice Stage exists).

That is, we ask when high-type Signalers earn less in the Signaling Stage by playing Punish/TFT compared to No Punish, in a game with no Partner Choice Stage. Recall that the payoff of Punish/ALLD for high-type Signalers is always lower than the payoff of Punish/TFT, so this condition is sufficient to exclude punishment in the absence of signaling.

From Eq. (5), we see that this condition is met when

$$h\left(\frac{b{w_H}^2 - c}{1 - {w_H}^4} + c\right) < c + k. \tag{13}$$

Thus, when the conditions specified in Eq. (11) and Eq. (13) are met, Punishment Signaling is an equilibrium, but no TPP will occur without Signaling. Note that the lower bound on $c + k$ given by Eq. (13) is guaranteed to be lower than the upper bound on $c + k$ given by Eq. (11) because $\frac{c}{b} < {w_H}^2$.

3.5. Example parameters

Here, we provide an example (visualized in Fig. 2) to illustrate possible conditions that could lead to the use of TPP as a signal. The following set of parameters (i) leads to a punishment signaling equilibrium (i.e. as specified by Eq. (1), $w_c w_s > \frac{c}{b}$ is satisfied, and Eq. (11) is also satisfied) (ii) does not lead to any punishment in the absence of signaling (i.e. Eq. (13) is satisfied):

- $h = .5$
- $w_H = .7$
- $w_L = .2$
- $b = 3$
- $c = 1$
- $k = 1$
- $j = 0$

3.6. Evolutionary dynamics

3.6.1. Methods

Here, we consider evolutionary dynamics to demonstrate that the Punishment Signaling equilibrium can be dynamically stable. We do so by simulating the transmission of strategies through an evolutionary process, which can be interpreted either as genetic evolution, or as social learning in which people copy the strategies of well-performing others. In both cases, strategies which earn higher payoffs are more likely to spread in the population, while lower-payoff strategies tend to die out. Novel strategies are introduced by mutation in the case of genetic evolution, or innovation and experimentation in the case of social learning. We use a frequency dependent Wright-Fisher model with an exponential payoff function to allow for negative payoffs with arbitrarily strong selection strength. We set a fixed proportion of agents, specified by $0 < h < 1$, to be high types, and the remaining proportion $1 - h$ to be low types. Agents only learn/inherent strategies from other agents of their own type. In any given interaction, each agent is equally likely to be the Signaler, Victim, or Chooser.

Each agent i has a strategy vector specifying how to behave when acting as Signaler, Victim, and Chooser. Specifically, an agent's Signaler strategy specifies whether to (i) play Punish+TFT, Punish+ALLD, or No Punish in the Signaling Stage and (ii) play TFT or ALLD in the Partner Choice Stage. An agent's Victim strategy specifies whether to (iii) play TFT or ALLD in the Signaling Stage. Finally, an agent's Chooser strategy specifies whether to (iv) play TFT or Reject when paired with a Signaler who punished, and (v) play TFT or Reject when paired with a Signaler who did not punish.

Agents in our model interact in a well-mixed population of constant size $N = 60$. Each generation, every agent plays in each of the three roles with every possible combination of other agents in the other two roles. The resulting payoff π_i is the sum of the expected payoffs for agent i over all roles and all combinations of other agents. We define agent i's fitness to be $f_i = e^{w\pi_i}$, where w is the intensity of selection; we set $w = 10$. Each generation, the entire population is updated (i.e. each agent updates his or her strategy). With probability $1 - u$, each agent picks an agent of his/her type proportional to fitness f, and takes on this agent's strategy; with probability u, a mutation occurs and instead the agent's strategy is replaced with a randomly selected strategy. Thus, u is the mutation rate; we set $u = .01$.

The purpose of our simulations is to demonstrate the stability of the Punishment Signaling equilibrium. To this end, we initialize the population such that 90% of agents begin with the Punishment Signaling equilibrium strategy and 10% of agents begin with random strategies. (Recall that the Punishment Signaling equilibrium strategy is for high types to play Punish+TFT as Signalers in the Signaling Stage; TFT as Signalers in the Partner Choice Stage; TFT as Victims in the Signaling Stage; and TFT when paired with a Signaler who punished, and Reject when paired with a Signaler who did not punish, as Choosers in the Partner Choice Stage. Low types instead play No Punish as Signalers in the Signaling Stage; ALLD as Signalers in the Partner Choice Stage; ALLD as Victims in the Signaling Stage; and Reject (regardless of if the Signaler punished) as Choosers in the Partner Choice Stage.)

We then record the strategies of all agents after every generation, and ask whether selection causes the population to converge on the Punishment Signaling equilibrium (indicating stability) or to diverge. We use the set of example parameters listed in Section 3.5 ($h = .5,\ w_H = .7,\ w_L = .2,\ b = 3,\ c = 1,\ j = 0$), while varying the value of k (the additional cost of engaging in TPP). For each set of parameters, we conduct 10 independent simulation runs, each with 1500 generations. We then investigate which values of k lead the population to converge on the Punishment Signaling equilibrium. In Fig. 3, we illustrate the parameter regions in which Punishment Signaling is Nash (between the dashed lines), and plot average strategies over the second half of generations for each parameter value (colored dotted lines).

3.6.2. Results

Broadly, Fig. 3 demonstrates that, as predicted, the Punishment Signaling equilibrium is dynamically stable in most of the region where it is Nash, and is not stable in the region where it is not Nash. Starting with Panel A, which shows Signaler behavior in the Signaling stage, we see that in the region where Punishment Signaling is Nash, high-type Signalers play Punish+TFT and low-type Signalers play No Punish, as specified by the Punishment Signaling strategy profile. In contrast, when Punishment Signaling is not Nash, both types play No Punish.

In Panel B, which shows Victim behavior in the Signaling stage, we see that in the region where Punishment Signaling is Nash, high-type Victims play TFT and low-type Victims play ALLD, as specified by the Punishment Signaling strategy profile. In contrast, when Punishment Signaling is not Nash, there is little selection on Victim strategies (because when Signalers do not punish, Victims do not get to play the Punishment PD, and thus Victim strategies do not affect Victim payoffs).

In Panel C, which shows Chooser behavior in the Partner Choice stage, we see that in the region where Punishment Signaling is Nash, high-type Choosers play TFT if the Signaler punished and Reject if the Signaler did not punish, and low-type Signalers always play Reject (regardless of Signaler punishment), as specified by the Punishment Signaling strategy profile. In contrast, when Punishment Signaling is not Nash, both types of Chooser play Reject if the Signaler did not punish, and typically also play Reject if the Signaler did punish. However, this latter parameter shows somewhat more noise (because when Signalers do not punish, Choosers do not encounter punishing Signalers, and thus their strategies in response to them do not affect their payoffs).

In Panel D, which shows Signaler behavior in the Partner Choice stage, we see that in the region where Punishment Signaling is Nash, high-type Signalers play TFT, as specified by the Punishment Signaling strategy profile. However, while the Punishment Signaling strategy profile also specifies that low-type Signalers play ALLD, we see that this strategy parameter shows some noise, reflecting that there is little selection on low-type Signaler strategies (because in equilibrium, Choosers reject Signalers who do not punish, and thus low-type Signalers do not get to play the Partner Choice PD, such that their Partner Choice PD strategies do not affect their payoffs). We note that for this reason, the Punishment Signaling strategy profile is not a *strict* Nash: low-type Signalers can neutrally deviate from ALLD to TFT in the Partner Choice stage. However, the Punishment Signaling equilibrium is still dynamically stable (i.e. all other strategy parameters in the Punishment Signaling equilibrium remain stable in our simulations) because this neutral deviation from low-type Signalers does not open the door for any other deviations, preventing indirect invasion by other equilibria. Looking at Signaler behavior in the Partner Choice stage in the region where Punishment Signaling is not Nash, we see that for this same reason, there is little selection on either high- or low-type Signaler strategies, because Choosers always reject both types of Signaler.

We note that the Punishment Signaling strategy profile is dynamically stable in most, but not all, of the region where it is Nash. When the additional cost of punishment k is close to the maximum possible value before high-type Signalers benefit from deviating to not punishing, the population converges on a different equilibrium in which both types of Signaler never punish and Choosers always reject. This likely reflects that mutation provides variation that can allow the “Pooling” equilibrium to invade as k approaches the limit of where Punishment Signaling is Nash, and thus the basin of attraction to the Punishment Signaling equilibrium becomes smaller. However, our results nonetheless demonstrate that there is a sizable region in which the Punishment Signaling equilibrium can be dynamically stable.

4. A model of dispersal as the origin of types

Thus far, we have assumed that the frequency of types (h) is fixed, and that agents cannot choose their type as part of their strategy. But where do types come from? Here we consider the possibility that types *can* evolve, and provide a simple model of dispersal that gives rise to the stable coexistence of types with low versus high continuation probabilities.

4.1. Dispersal game

We present a scenario in which agents are engaged in a “dispersal game” as well as the signaling game described above. In this game, agents can either stay in their local environment, or leave (i.e. disperse). Leaving confers a fixed (expected) benefit α, while staying confers a frequency-depend benefit that is decreasing in the fraction of the population that stays (because of competition over limited resources). Specifically, the payoff of staying is defined as

$$\frac{\beta}{af+1} \tag{14}$$

where f is the fraction of the population that stays, and a is a parameter controlling how frequency-dependent the payoff of staying is (with higher values of a meaning the payoff of staying declines more rapidly when others stay).

We assume that agents can choose between two possible strategies: staying with probability w_L and staying with probability w_H, where w_L and w_H are the same as the continuation probabilities for low and high types from our main model. In other words, agents can choose between being low or high types by setting their probability of staying in the dispersal game to be low or high. Thus, in a population where the proportion of high types is h, low types earn

$$(1 - w_\alpha)\alpha + w_\alpha \frac{\beta}{a(hw_H + (1-h)w_\alpha) + 1} \tag{15}$$

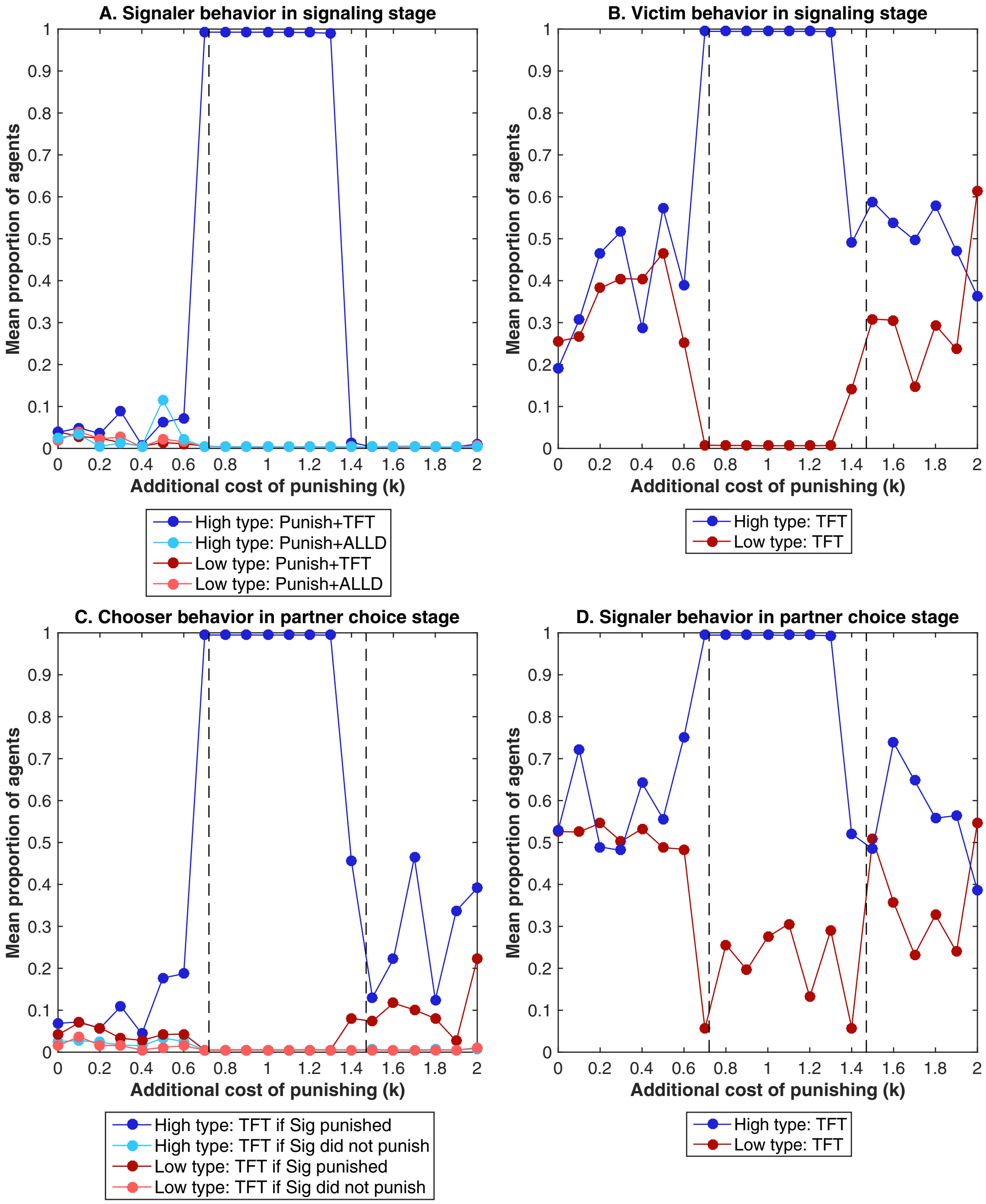


Fig. 3. Shown is the region in which Punishment Signaling is Nash (between the dashed vertical lines), and the average proportion of high-type (blue dotted lines) and low-type (red dotted lines) agents (over the second half of generations in all simulation runs) playing the specified strategies, as a function of k (the additional cost of TPP). A) Signaler behavior in signaling stage. Shown is the proportion of agents playing Punish+TFT (dark blue and dark red) and Punish+ALLD (light blue and light red) (with all remaining agents playing No Punish) as Signalers in the Signaling Stage. B) Victim behavior in signaling stage. Shown is the proportion of agents playing TFT (dark blue and dark red) (with all remaining agents playing ALLD) as Victims in the Signaling Stage. C) Chooser behavior in partner choice stage. Shown is the proportion of agents playing TFT (with all remaining agents playing Reject) when paired with Signalers who did punish (dark blue and dark red) versus did not punish (light blue and light red), as Choosers in the Partner Choice Stage. D) Signaler behavior in partner choice stage. Shown is the proportion of agents playing TFT (dark blue and dark red) (with all remaining agents playing ALLD) as Signalers in the Partner Choice Stage. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and high types earn

$$(1 - w_H)\alpha + w_H \frac{\beta}{a(hw_H + (1-h)w_\alpha) + 1}. \qquad (16)$$

4.2. Conditions for stable types

Based on these payoffs, we can derive the conditions for co-existence of low and high types. This occurs when the payoff to high and low types are identical, when simultaneously considering payoffs summed across the dispersal game and the signaling game, which are as follows.

(a) Payoff for low types

i. In the dispersal game, from Eq. (15): $(1 - w_\alpha)\alpha + w_\alpha \frac{\beta}{a(hw_H+(1-h)w_\alpha)+1}$

   ii. In the signaling game, as the Chooser (playing Reject), from Eq. (1): 0
   iii. In the signaling game, as the Signaler (playing No Punish), from Eq. (1): 0

(b) Payoff for high types
   i. In the dispersal game, from Eq. (16): $(1-w_H)\alpha + w_H\frac{\beta}{a(hw_H+(1-h)w_\alpha)+1}$
   ii. In the signaling game, as the Chooser (playing TFT iff Signaler punishes, and thus is high-type TFT player), from Eq. (2): $h\frac{bw_H{}^2-c}{1-w_H{}^4}$
   iii. In the signaling game, as the Signaler (playing Punish/TFT), from Eq. (8): $-k+h\left(\frac{(b-c)(1+w_H{}^2)}{1-w_H{}^4}\right)-(1-h)c$

Assuming the players are equally likely to be in the Chooser and Signaler roles when playing the signaling game, this yields a total payoff for low types of

$$\pi_\alpha = (1-w_\alpha)\alpha + w_\alpha\frac{\beta}{a(hw_H+(1-h)w_\alpha)+1} \tag{17}$$

and a total payoff for high types of

$$\pi_H = (1-w_H)\alpha + w_H\frac{\beta}{a(hw_H+(1-h)w_\alpha)+1} + \frac{1}{2}\frac{h(bw_H{}^2-c)}{1-w_H{}^4} + \frac{1}{2}\left(-k+\frac{h(b-c)(1+w_H{}^2)}{1-w_H{}^4}-(1-h)c\right) \tag{18}$$

Thus, there is stable coexistence between low and high types when

$$\pi_\alpha = \pi_H \tag{19}$$

as neither type has an incentive to switch.

4.2.1. Example parameters

Here, we provide an example to illustrate possible conditions that could lead to stable coexistance of types. Specifically, we provide sample dispersal game parameters that equalize payoffs to each type, in the context of the example signaling game parameters provided in Section 3.5 (reproduced below for reference). The following set of parameters (i) leads to a stable TPP signaling equilibrium (i.e. Eq. (11) is satisfied) (ii) does not lead to any punishment in the absence of signaling (i.e. Eq. (13) is satisfied) and gives equal aggregate payoffs to both low and high types (i.e. Eq. (19) is satisfied), such that there is no incentive for deviation in any part of the model.

- In the dispersal game
  - $\alpha = 1$
  - $\beta = 5$
  - $a = 10740/233$
- In the signaling game
  - $h = .5$
  - $w_H = .7$
  - $w_\alpha = .2$
  - $b = 3$
  - $c = 1$
  - $k = 1$
  - $j = 0$

5. Imperfect signals

In the concrete model that we have described thus far, third-party punishment is a *perfect* signal of type: Signalers have the same probability of future interaction in the Signaling Stage as the Partner Choice Stage, such that individuals who find investing in reciprocity via TPP in the Signaling Stage relatively beneficial *always* also find investing in reciprocity via cooperation in the Partner Choice Stage relatively beneficial. However, in JHBR, TPP is only an *imperfect* signal of type, because both trustworthy and exploitative Signalers sometimes experience small and large costs of punishing. (To put this in the context of the current model, JHBR considers the situation in which Signalers with high vs low probabilities of future interaction in the Partner Choice Stage both sometimes experience high vs low probabilities of future interaction with the Victim in the Signaling Stage). Furthermore, in JHBR, there are two signals of the same underlying trait that vary in their informativeness (punishment and helping).

These features could be incorporated into the concrete model presented here by specifying that in the Signaling Stage, a Signaler's probability of experiencing a "high" probability of future interaction is only *probabilistically* based on her type from the Partner Choice Stage. To this end, we could specify that individual Signalers experience different probabilities of future interaction in different Signaling Stage "signaling PDs" (sometimes punishment PDs, sometimes helping PDs) but that high-type Signalers (i.e. Signalers with a high probability of future interaction in the Partner Choice Stage) are *more likely* than low-type Signalers to experience high probabilities of future interaction in both signaling PDs. Furthermore, if helping was to be a more informative signal of punishment, this would be more true for the helping PD than the punishment PD.

The interpretation of this would be that incentives to signal in a signaling context are probabilistically informative of an individual's incentives to cooperate in a partner choice context (specifically, their probability of future interaction in the Partner Choice Stage). Then, the informativeness of a particular signal would depend on how correlated those incentives were (that is, how similar is the probability of future interaction in the signaling context and in the partner choice context?).

In such a model, a Signaler's "type" in each signaling PD (i.e. probability of future interaction in that PD) could be determined probabilistically as a function of her partner choice type, as follows:

	Signaler type in Partner Choice Stage	
Probability of being a...	High	Low
High type in Punishment PD	c_1	c_2
Low type in Punishment PD	$1-c_1$	$1-c_2$
High type in Helping PD	c_3	c_4
Low type in Helping PD	$1-c_3$	$1-c_4$

And then, as in JHBR, the informativeness of punishment would be specified by $I_{SP}=\frac{c_1}{c_2}$ and $I_{LP}=\frac{1-c_2}{1-c_1}$ and the informativeness of helping would be specified by $I_{SH}=\frac{c_3}{c_4}$ and $I_{LH}=\frac{1-c_4}{1-c_3}$. These values, together with the mapping given in Eq. (4), could then be plugged into the calculations given in JHBR to derive the conditions for the various equilibria described therein.

6. Conclusion

TPP is a critical part of human social behavior (e.g. Boyd et al., 2003; Fehr and Fischbacher, 2004; Henrich et al., 2006; Mathew and Boyd, 2011; McAuliffe et al., 2015), but a key question is why individuals are willing to pay costs (Balafoutas et al., 2014; Boyd et al., 2010) to punish. Here, we have focused specifically on the question of why people punish in cases where the costs of punishing (e.g. retaliation) outweigh any benefits they will derive from deterring future selfish behavior towards others. Building on a large body of research implicating a role of reputation in TPP (Barclay, 2006; Boyd and Richerson, 1992; Horita, 2010; Kurzban et al., 2007; Nelissen, 2008; Panchanathan and Boyd, 2004; Raihani and Bshary, 2015a; 2015b), we have followed up on a recent model of TPP as a costly signal of trustworthiness

(Jordan et al., 2016). We have complemented this abstract model (in which TPP is argued to generally signal exposure to mechanisms incentivizing cooperation) with a more concrete model demonstrating how this could work in the context of repeated interactions and direct reciprocity.

In this more concrete model, trustworthy types are individuals who are likely to be present in the future, and thus who benefit from investing in reciprocity. Punishment is less costly for trustworthy types because they are more likely to receive future reciprocity from victims on whose behalf they punish, and as such third-party punishment can function as a costly signaling of trustworthiness. We have described the modeling framework in detail, and shown how costly signaling can allow a stable equilibrium with third-party punishment in parameter regions where no punishment would occur without signaling. We have also shown how types could originate from a simple dispersal game with frequency-dependent payoffs, and how our model could incorporate multiple signals that vary in their informativeness.

We note that in both the Partner Choice PD (Partner Choice Stage) and the Punishment PD (Signaling Stage), we use a restricted strategy set, in which the only possible repeated PD strategies are TFT and ALLD; agents are not allowed, for example, to play ALLC or more complex conditional strategies (e.g. Tit for Two Tats). Our goal in selecting this strategy set is to create conditions that straightforwardly favor a TFT equilibrium when the relevant continuation probability is high enough, relative to the c/b, and then to use this simple model to illustrate the general point that exposure to a common mechanism for TPP and cooperation (in this case, direct reciprocity supported by a high continuation probability) can allow TPP to function as a costly signal of trustworthiness.

However, as is the case of many direct reciprocity models, allowing for other strategies can undermine the stability of TFT as an equilibrium. In particular, including ALLC as a strategy can lead to cylces of indirect invasion: in equilibrium, deviating from TFT to ALLC is neutral, but opens the door for ALLD to invade ALLC (Imhof et al., 2005). While this indirect invasion reduces the steady state level of cooperation, the population can still spend a great deal of time playing TFT because stochastic dynamics allow TFT to re-invade ALLD. Furthermore, various mechanisms may function to prevent indirect invasion by ALLC, such as a small amount of assortment (Van Veelen et al., 2012), the introduction of errors and the use of Win-Stay-Lose-Shift (Nowak and Sigmund, 1993), or mutation maintaining a background fraction of ALLD. While we chose, for simplicity, to leave these strategies, mechanisms, and resulting dynamics out of our model, they are likely to underlie direct reciprocity-based equilibria (like the Punishment Signaling equilibrium discussed here) the real world, and incorporating such complexities into our framework is a promising direction for future work.

While our model has focused on direct reciprocity as the common mechanism incentivizing trustworthiness and TPP, the same logic can also apply to other mechanisms that support both cooperation and punishment. For example, you could easily move from direct reciprocity to indirect reciprocity or institutional reward (whereby cooperation and punishment are rewarded not by the recipient of the cooperation / the victim of the punished transgression, but other observers or institutions or leaders). In this example, you can think of w_L and w_H as the probabilities not of future interaction with the recipient / victim, but rather observers, institutions, or leaders. Future theoretical work should describe in detail such alternative mechanisms, as well as alternative methods for generating stable coexistence of types; and future empirical work should evaluate the extent to which incentives to cooperate are correlated with deterrence-based benefits of TPP outside the laboratory.

We also note that the signaling mechanism modeled here may also help explain cases of TPP that are totally anonymous, or otherwise cannot actually elicit signaling benefits. Our model provides an ultimate explanation for why TPP may be advantageous to individuals; at a proximate level, however, third-party punishers may not be consciously seeking to signal their trustworthiness, and may instead be driven by moral outrage (Fehr and Fischbacher, 2004; Jordan et al., 2015) or other emotions. Because TPP may be based on internalized emotions, ideologies, or social heuristics (Bear and Rand, 2016; Rand, 2016; Bear, Kagan, and Rand, 2017), rather than explicit strategic calculations, people may still be motivated to punish even when their punishment cannot function as a signal. In other words, punishment motives that are ultimately explained by signaling may ”spill over” (Peysakhovich and Rand, 2015) to anonymous settings, or contexts where signaling trustworthiness is not actually beneficial (e.g. when people observing the punishment will not have the opportunity to invest in trusting the punisher).

Appendix A

In the main text, we focused on the case where $w_L w_H < \frac{c}{b} < {w_H}^2$, such that TFT is only an equilibrium in the Partner Choice Stage when both players are high types. When $\frac{c}{b} < {w_L}^2$, TFT is an equilibrium for all Signaler-Chooser pairs, and no signaling will occur. And when ${w_H}^2 < \frac{c}{b}$, TFT is never any equilibrium for any Signaler-Chooser pairs, and signaling will also never occur. But what happens when ${w_L}^2 < \frac{c}{b} < w_L w_H$?

A1. Additional Punishment Signaling strategy profile

When ${w_L}^2 < \frac{c}{b} < w_L w_H$, TFT is an equilibrium in the Partner Choice stage as long as at least one of the Chooser and Signaler is a high type (unlike the case we focus on in the main text, $w_L w_H < \frac{c}{b} < {w_H}^2$, where TFT is only an equilibrium when both players are high types). Here, we demonstrate the conditions for a different Punishment Signaling strategy profile to be an equilibrium when ${w_L}^2 < \frac{c}{b} < w_L w_H$. In this equilibrium, high-type Signalers use TPP to signal to *low-type* Choosers, rather than *high-type* Choosers. The strategy profile is defined as follows. In the Signaling Stage, (A) high-type Signalers play Punish/TFT; and (B) low-type Signalers play No Punish. In the Partner Choice stage, (C) high-type Choosers play TFT; (D) high-type Signalers play TFT; (E) low-type Choosers play TFT if and only if the Signaler punished; and (F) low-type Signalers play TFT. We now derive the conditions required for there to be no incentive to deviate from each of these six features of the strategy profile, under the assumption that ${w_L}^2 < \frac{c}{b} < w_L w_H$.

A2. Signaling stage results

A2.1. (A) High-type Signalers play Punish/TFT

Starting with the signaling stage, we begin by deriving the conditions required for (A) - that is, for high-type Signalers to not improve their payoff by deviating from playing Punish/TFT (given that on the equilibrium path, high-type Signalers must punish in order to elicit TFT from low-type Choosers in the partner choice PD).

First, we calculate the total payoff to high-type Signalers in equilibrium. Calculating their payoff in the Signaling Stage requires specifying the Victim's punishment PD strategy. On the equilibrium path, if the Signaler opens the punishment PD by punishing in the punishment phase, the Victim can infer that the Signaler is a high type and therefore that the Signaler will play TFT in the cooperation phase (because on equilibrium path, low-type Signalers do not punish). It is payoff-maximizing for the Victim to respond to a Signaler who plays Punish/TFT with TFT when $\frac{c}{b} < w_v w_s$, which is always satisfied when the Signaler is a high type (because $\frac{c}{b} < w_L w_H$). Thus, the Victim will always respond to

Punish/TFT with TFT, regardless of his own type. However, the payoff to the Signaler from interacting with the Victim will vary, depending on the Victim's type and resulting value of w_v.

A high-type Signaler's expected payoff in the Signaling Stage from playing Punish/TFT is therefore

$$-k+h\frac{b{w_H}^2-c}{1-{w_H}^4}+(1-h)\frac{bw_Hw_L-c}{1-(w_Hw_L)^2}. \tag{A.1}$$

Then, in the Partner Choice Stage, high-type Signalers will receive $\frac{b-{w_H}^2c}{1-{w_H}^4}$ with probability h (when the Signaler is paired with a high-type Chooser, given that on the equilibrium path high-type Choosers and Signalers both always play TFT); and $\frac{b-w_Hw_Lc}{1-(w_Hw_L)^2}$ with probability $1-h$ (when the Signaler is paired with a low-type Chooser, given that on the equilibrium path low-type Choosers play TFT if the Signaler punished, and high-type Signalers always play TFT). Thus, on the equilibrium path, across both stages a high-type Signaler playing Punish/TFT earns a total payoff of

$$-k+h\frac{(b-c)(1+{w_H}^2)}{1-{w_H}^4}+(1-h)\frac{(b-c)(1+w_Hw_L)}{1-(w_Hw_L)^2}. \tag{A.2}$$

For this behavior to be in equilibrium, the high-type Signaler must not be able to improve her payoff by switching to either Punish/ALLD or No Punish.

If the high-type Signaler deviates by playing Punish/ALLD, she earns an identical payoff in the Partner Choice Stage (because the Chooser's behavior in Partner Choice Stage is conditional only on whether the Signaler punishes). However, her expected payoff in the Signaling Stage becomes $-k+h(b{w_H}^2-c)+(1-h)(bw_Hw_L-c)$, which is always lower than the Signaling Stage payoff of playing Punish/TFT (because $\frac{c}{b}<w_Hw_L$).

If the high-type Signaler deviates by playing No Punish, she earns zero payoff in the Signaling Stage. In the Partner Choice stage, she earns $\frac{b-{w_H}^2c}{1-{w_H}^4}$ with probability h (when the Signaler is paired with a high-type Chooser, given that on the equilibrium path high-type Choosers always play TFT); and 0 with probability $1-h$ (when the Signaler is paired with a low-type Chooser, given that on the equilibrium path low-type Choosers play Reject if the Signaler did not punish). Asking when Eq. (A.2) is greater than this payoff from deviating shows that a high-type Signaler cannot improve her payoff by deviating from Punish/TFT when

$$k<h\frac{b{w_H}^2-c}{1-{w_H}^4}+(1-h)\frac{(b-c)(1+w_Hw_L)}{1-(w_Hw_L)^2}. \tag{A.3}$$

A2.2. (B) Low-type Signalers play No Punish

Next, we derive the conditions required for (B) - that is, for low-type Signalers to not improve their payoffs by deviating from playing No Punish in the Signaling Stage. First, we calculate the total payoff to low-type Signalers in equilibrium. In the Signaling Stage, they do not punish and receive 0. In the Partner Choice Stage, they receive $\frac{b-w_Hw_Lc}{1-(w_Hw_L)^2}$ with probability h (when they meet a high-type Chooser, who plays TFT, to which they respond with TFT) and 0 with probability $1-h$ (when they meet a low-type Chooser, who plays Reject). Thus, their total payoff is $h\frac{b-w_Hw_Lc}{1-(w_Hw_L)^2}$.

If a low-type Signaler deviates to Punishing in the Signaling Stage, the Victim will always respond by playing TFT (because on equilibrium path, only high-type Signalers punish). If the low-type Signaler deviates to playing Punish/ALLD, her expected payoff in the Signaling Stage thus becomes $-k+h(bw_Hw_L-c)+(1-h)(b{w_L}^2-c)$. If she instead deviates to playing Punish/TFT, her expected payoff in the Signaling stage becomes $-k+h\frac{bw_Hw_L-c}{1-(w_Hw_L)^2}+(1-h)\frac{b{w_L}^2-c}{1-{w_L}^4}$. Because ${w_L}^2<\frac{c}{b}<w_Hw_L$, and thus it is better for a low-type Signaler to play Punish/TFT than Punish/ALLD when paired with a high-type Victim but Punish/ALLD than Punish/TFT when paired with a low-type Victim, the deviation that earns her a higher payoff depends on the value of h (i.e. the probability that she is paired with a high-type Victim).

Following either of these deviations, in the Partner Choice Stage, the Chooser will always play TFT, given that on the equilibrium path high-type Choosers always play TFT and low-type Choosers play TFT if and only if the Signaler punished. On equilibrium path, low-type Signalers always respond to TFT with TFT (because only high-type Choosers play TFT when paired with Signalers who have not punished, and low-type Signalers find it payoff-maximizing to engage in reciprocal cooperation with high-type Choosers because $\frac{c}{b}<w_Hw_L$). If the low-type Signaler plays TFT in the Partner Choice stage after punishing, she will receive $\frac{b-w_Hw_Lc}{1-(w_Hw_L)^2}$ with probability h and $\frac{b-{w_L}^2c}{1-{w_L}^4}$ with probability $1-h$. If the low-type Signaler instead plays ALLD in the Partner Choice stage, she will receive b. Again, because ${w_L}^2<\frac{c}{b}<w_Hw_L$, and thus it is better to play TFT than ALLD when paired with a high-type Chooser but ALLD than TFT when paired with a low-type Chooser, the deviation that earns low-type Signalers a higher payoff depends on the value of h (i.e. the probability that she is paired with a high-type Chooser). Furthermore, whenever h is high enough that a low-type Signaler earns a higher payoff from playing Punish/TFT than Punish/ALLD in the Signaling Stage, she will earn a higher payoff from subsequently playing TFT than ALLD in the Partner Choice Stage (and vice versa).

Thus, if the low-type Signaler deviates to playing Punish/ALLD in the Signaling Stage and then ALLD in the Partner Choice Stage, she will earn a total payoff across both stages of $-k+h(bw_Hw_L-c)+(1-h)(b{w_L}^2-c)+b$. And if she deviates to playing Punish/TFT in the Signaling Stage and then TFT in the Partner Choice Stage, she will earn a total payoff of $-k+h\frac{(b-c)(1+w_Hw_L)}{1-(w_Hw_L)^2}+(1-h)\frac{(b-c)(1+{w_L}^2)}{1-{w_L}^4}$. Both of these deviations are negative when

$$k>h\left(bw_Hw_L-c-\frac{b-w_Hw_Lc}{1-(w_Hw_L)^2}\right)+(1-h)\left(b{w_L}^2-c\right)+b \tag{A.4}$$

and

$$k>h\frac{bw_Hw_L-c}{1-(w_Hw_L)^2}+(1-h)\frac{(b-c)(1+{w_L}^2)}{1-{w_L}^4} \tag{A.5}$$

A3. Partner choice stage results

A3.1. (C) High-type Choosers play TFT

We now move to the partner choice stage. On equilibrium path, both high-type and low-type Signalers play TFT in the partner choice stage. Choosers thus cannot benefit from deviating from themselves playing TFT when $\frac{c}{b}<w_cw_s$ which is always true for high-type Choosers because $\frac{c}{b}<w_Lw_H$. Thus high-type Choosers cannot benefit from deviating from TFT.

A3.2. (D) High-type Signalers play TFT

The same logic applies to high-type Signalers, who also cannot benefit from deviating from TFT when $\frac{c}{b}<w_cw_s$. This is always true for high-type Signalers because $\frac{c}{b}<w_Lw_H$, so high-type Signalers cannot benefit from deviating from TFT.

A3.3. (E) Low-type Choosers play TFT if and only if the Signaler punished

On the equilibrium path, if a Signaler punished, she is a high type, and if a Signaler did not punish, she is a low type. Because ${w_L}^2<\frac{c}{b}<w_Lw_H$, low-type Choosers paired with high-type Signalers earn a higher payoff from sticking with TFT than from

switching to Reject; and low-type Choosers paired with low-type Signalers earn a higher payoff from sticking with Reject than from switching to TFT. Thus, low-type Choosers cannot benefit from deviating from playing TFT if and only if the Signaler punished.

A3.4. (D) Low-type Signalers play TFT

On the equilibrium path, if a Chooser plays TFT when paired with a low-type Signaler (who did not punish), he is a high type. Because $\frac{c}{b} < w_L w_H$, low-type Signalers who are paired with high-type Choosers earn a higher payoff from sticking with TFT than from switching to ALLD. Thus, low-type Signalers cannot benefit from deviating from TFT.

A4. Conditions for additional Punishment Signaling equilibrium

Taken together, this additional Punishment Signaling strategy profile, in which high-type Signalers punish to signal their type to low-type Choosers, is an equilibrium when ${w_L}^2 < \frac{c}{b} < w_H w_L$ is met, and

$$k < h\frac{b{w_H}^2 - c}{1 - {w_H}^4} + (1-h)\frac{(b-c)(1+w_H w_L)}{1-(w_H w_L)^2} \tag{A.6a}$$

$$h(bw_H w_L - c - \frac{b - w_H w_L c}{1-(w_H w_L)^2}) + (1-h)(b{w_L}^2 - c) + b < k \tag{A.6b}$$

$$h\frac{bw_H w_L - c}{1-(w_H w_L)^2} + (1-h)\frac{(b-c)(1+{w_L}^2)}{1-{w_L}^4} < k \tag{A.6c}$$

are all satisfied.

A5. Conditions for no punishment in the absence of signaling

Next, we ask when, under this additional Punishment Signaling equilibrium, no punishment will occur *without* signaling. In other words, we ask when high-type Signalers earn less in the Signaling Stage by playing Punish/TFT compared to No Punish, in a game with no Partner Choice Stage. Recall that the payoff of Punish/ALLD for high-type Signalers is always lower than the payoff of Punish/TFT, so this condition is sufficient to exclude punishment in the absence of signaling. This occurs when:

$$h\frac{b{w_H}^2 - c}{1 - {w_H}^4} + (1-h)\frac{bw_H w_L - c}{1-(w_H w_L)^2} < k \tag{A.7}$$

A6. Example parameters

Here, we provide an example to illustrate possible conditions that could lead to the use of TPP as a signal under this additional Punishment Signaling equilibrium. The following set of parameters (i) leads to a stable TPP signaling equilibrium (i.e. ${w_L}^2 < \frac{c}{b} < w_H w_L$ is met, and all parts of Eq. (A.6) are also satisfied) and (ii) does not lead to any punishment in the absence of signaling (i.e. Eq. (A.7) is satisfied):

- $h = .5$
- $w_H = .7$
- $w_L = .5$
- $b = 3$
- $c = 1$
- $k = 1.5$
- $j = 0$

References

Balafoutas, L., Grechenig, K., Nikiforakis, N., 2014. Third-party punishment and counter-punishment in one-shot interactions. Econ. Lett. 122, 308–310.

Barclay, P., 2006. Reputational benefits for altruistic punishment. Evol. Human Behav. 27, 325–344.

Bear, A., Kagan, A., Rand, D.G., 2017. Co-evolution of cooperation and cognition: the impact of imperfect deliberation and context-sensitive intuition. In: Proc. R. Soc. B, Vol. 284, N0. 1851. The Royal Society, p. 20162326.

Bear, A., Rand, D.G., 2016. Intuition, deliberation, and the evolution of cooperation. Proc. Natl. Acad. Sci. 113, 936–941.

Bliege Bird, R., Smith, E., Bird, D.W., 2001. The hunting handicap: costly signaling in human foraging strategies. Behav. Ecol. Sociobiol. 50 (1), 9–19.

Boone, J.L., 1998. The evolution of magnanimity. Human Nat. 9 (1), 1–21.

Boyd, R., Gintis, H., Bowles, S., 2010. Coordinated punishment of defectors sustains cooperation and can proliferate when rare. Science 328, 617–620.

Boyd, R., Gintis, H., Bowles, S., Richerson, P.J., 2003. The evolution of altruistic punishment. Proc. Natl. Acad. Sci. 100, 3531–3535.

Boyd, R., Richerson, P.J., 1992. Punishment allows the evolution of cooperation (or anything else) in sizeable groups. Ethol. Sociobiol. 13, 171–195.

Charness, G., Cobo-Reyes, R., Jimenez, N., 2008. An investment game with third–party intervention. J. Econ. Behav. Org. 68, 18–28.

Fehr, E., Fischbacher, U., 2004. Third-party punishment and social norms. Evol. Human Behav. 25, 63–87.

Fowler, J.H., 2005. Altruistic punishment and the origin of cooperation. Proc. Natl. Acad. Sci. USA 102, 7047–7049.

Garcia, J., Traulsen, A., 2012. Leaving the loners alone: evolution of cooperation in the presence of antisocial punishment. J. Theor. Biol. 307, 168–173.

Gintis, H., Smith, E.A., Bowles, S., 2001. Costly signaling and cooperation. J. Theor. Biol. 213 (1), 103–119.

Hamlin, J.K., Wynn, K., Bloom, P., Mahajan, N., 2011. How infants and toddlers react to antisocial others. Proc. Natl. Acad. Sci. 108, 19931–19936.

Hauert, C., Traulsen, A., Brandt, H., Nowak, M.A., Sigmund, K., 2007. Via freedom to coercion: the emergence of costly punishment. Science 316, 1905–1907.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., Cardenas, J.C., Gurven, M., Gwako, E., Henrich, N., Lesorogol, C., Marlowe, F.W., Tracer, D., Ziker, J., 2006. Costly punishment across human societies. Science 312, 1767–1770.

Hilbe, C., Traulsen, A., 2012. Emergence of responsible sanctions without second order free riders, antisocial punishment or spite. Sci. Rep. 2, 458.

Horita, Y., 2010. Punishers may be chosen as providers but not as recipients. Lett. Evol. Behav. Sci. 1, 6–9.

Imhof, L.A., Fudenberg, D., Nowak, M.A., 2005. Evolutionary cycles of cooperation and defection. Proc. Natl. Acad. Sci. 102 (31), 10797–10800.

Jordan, J., Peysakhovich, A., Rand, D.G., 2015. Why we cooperate. Moral Brain: A Multidiscip. Perspective 87.

Jordan, J.J., Hoffman, M., Bloom, P., Rand, D.G., 2016. Third-party punishment as a costly signal of trustworthiness. Nature 530 (7591), 473–476.

Jordan, J.J., McAuliffe, K., Rand, D.G., 2015. The effects of endowment size and strategy method on third party punishment. Exp. Econ.

Jordan, J.J., McAuliffe, K., Warneken, F., 2014. Development of in-group favoritism in childrens third-party punishment of selfishness. Proc. Natl. Acad. Sci. 111, 12710–12715.

Kurzban, R., DeScioli, P., O'Brien, E., 2007. Audience effects on moralistic punishment. Evol. Human Behav. 28, 75–84.

Mathew, S., Boyd, R., 2011. Punishment sustains large-scale cooperation in prestate warfare. Proc. Natl. Acad. Sci. 108, 11375–11380.

McAuliffe, K., Jordan, J.J., Warneken, F., 2015. Costly third-party punishment in young children. Cognition 134, 1–10.

Nakamaru, M., Iwasa, Y., 2006. The coevolution of altruism and punishment: role of the selfish punisher. J. Theor. Biol. 240, 475–488.

Nelissen, R.M.A., 2008. The price you pay: cost-dependent reputation effects of altruistic punishment. Evol. Human Behav. 29, 242–248.

Nowak, M., Sigmund, K., 1993. A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game. Nature 364 (6432), 56.

Nowak, M.A., 2006. Five rules for the evolution of cooperation. Science 314, 1560–1563.

Nowak, M.A., Page, K.M., Sigmund, K., 2000. Fairness versus reason in the ultimatum game. Science 289, 1773–1775.

Panchanathan, K., Boyd, R., 2004. Indirect reciprocity can stabilize cooperation without the second-order free rider problem. Nature 432, 499–502.

Peysakhovich, A., Rand, D.G., 2015. Habits of virtue: creating norms of cooperation and defection in the laboratory. Manage. Sci. 62, 631–647.

Raihani, N.J., Bshary, R., 2015a. The reputation of punishers. Trends in Ecol. Evol. 30, 98–103.

Raihani, N.J., Bshary, R., 2015b. Third party punishers are rewarded but third party helpers even more so. Evolution 69, 993–1003.

Rand, D.G., 2016. Cooperation, fast and slow: meta-analytic evidence for a theory of social heuristics and self-interested deliberation. Psychol. Sci.

Rand, D.G., Armao IV, J.J., Nakamaru, M., Ohtsuki, H., 2010. Anti-social punishment can prevent the co-evolution of punishment and cooperation. J. Theor. Biol. 265, 624–632.

Rand, D.G., Nowak, M.A., 2011. The evolution of antisocial punishment in optional public goods games. Nat. Commun. 2, 434.

Rand, D.G., Tarnita, C.E., Ohtsuki, H., Nowak, M.A., 2013. Evolution of fairness in the one-shot anonymous ultimatum game. Proc. Natl. Acad. Sci. 110, 2581–2586.

Roberts, G., 1998. Competitive altruism: from reciprocity to the handicap principle. Proc. R. Soc. Lond. B: Biol. Sci. 265 (1394), 427–431.

Roberts, G., 2013. When punishment pays. PloS One 8 (e5), 7378.

Rockenbach, B., Milinski, M., 2006. The efficient interaction of indirect reciprocity and costly punishment. Nature 444, 718–723.

Roff, D.A., 1975. Population stability and the evolution of dispersal in a heterogeneous environment. Oecologia 19, 217–237.

dos Santos, M., Rankin, D.J., Wedekind, C., 2010. The evolution of punishment through reputation. Proc. R. Soc. Lond. B: Biol. Sci.

Smith, E.A., Bliege Bird, R., 2000. Turtle hunting and tombstone opening: public generosity as costly signaling. Evol. Human Behav. 21 (4), 245–261.

Smith, E.A., Bliege Bird, R., 2005. Costly signaling and cooperative behavior. In: Moral Sentiments and Material Interests: The Foundations of Cooperation in Economic Life, pp. 115–148.

Spence, M., 1973. Job market signaling. Quart. J. Econ. 87 (3), 355–374.

Tarnita, C.E., 2015. Fairness and trust in structured populations. Games 6, 214–230.

Trivers, R.L., 1971. The evolution of reciprocal altruism. Quart. Rev. Biol. 35–57.

Van Veelen, M., Garca, J., Rand, D.G., Nowak, M.A., 2012. Direct reciprocity in structured populations. Proc. Natl. Acad. Sci. 109 (25), 9929–9934.

Wright, J., 1999. Altruism as a signal: Zahavi's alternative to kin selection and reciprocity. J. Avian Biol. 108–115.

Zahavi, A., 1975. Mate selectiona selection for a handicap. J. Theor. Biol. 53, 205–214.

Zahavi, A., 1977. Reliability in communication systems and the evolution of altruism. In: Evolutionary Ecology. Macmillan Education UK, pp. 253–259.

Zahavi, A., 1995. Altruism as a handicap: the limitations of kin selection and reciprocity. J. Avian Biol. 26 (1), 1–3.