nature human behaviour

ARTICLES

https://doi.org/10.1038/s41562-020-01008-1

Arbitration supports reciprocity when there are frequent perception errors

Robert Boyd[1,2] and Sarah Mathew[1,2] ✉

Reciprocity is undermined by perception errors, mistakes that cause disagreement between interacting individuals about past behaviour. Strategies such as win–stay–lose–shift and generous tit-for-tat can re-establish cooperation following a perception error, but only when errors arise infrequently. We introduce arbitration tit-for-tat (ATFT), a strategy that uses third-party arbitration to align players' beliefs about what transpired when they disagree. We show that, when arbitration is moderately accurate, ATFT is a strong subgame-perfect equilibrium and is evolutionarily stable against a range of strategies that defect, cooperate, ignore arbitration or invoke arbitration unnecessarily. ATFT can persist when perception errors are frequent, arbitration is costly or arbitration is biased. The need for third parties to resolve perception errors could explain why reciprocity is rare in other animals despite opportunities for repeated interactions and why human reciprocity is embedded within culturally transmitted moral norms in which community monitoring plays a role.

The evolution of reciprocity[1,2] can be sensitive to behavioural errors because a mistaken defection by one individual can motivate one's partner to defect, setting off a sequence of reprisals that undermine cooperation[3–7]. This need not occur if individuals know that they have erred, an 'execution error', because the individual who mistakenly defects can accept their partner's retaliation without further defection and this act of contrition restores mutual cooperation[3–5]. However, individuals may often be unaware that they have erred. Such perception errors[5–7] pose a greater challenge because partners hold different beliefs about what transpired, which often leads to the collapse of cooperation. Strategies such as win–stay–lose–shift (WSLS)[8–10] and generous tit-for-tat (GTFT)[11–13] can resolve this problem when errors are rare, but have not been shown to sustain cooperation when there are frequent perception errors. When the right mix of invading strategies is present, strategies such as GTFT and WSLS are vulnerable to indirect invasion[14,15]. For example, GTFT can be invaded by always cooperate (ALLC), the strategy that always cooperates, which can then be invaded by defecting strategies and this can lead to cycles or the collapse of cooperation. Such indirect invasion need not occur when individuals know they have erred[4].

While perception error rates have never been measured, there are compelling reasons to think that perception errors arise frequently in naturalistic interactions. In most models the costs and benefits to the recipient of helping are fixed. In the real world the costs and benefits vary, occur in different currencies and depend on many contingencies, especially the states of an individual that are not known to others[16,17]. Individuals often need to infer whether their partner has cooperated based on incomplete information about their partner's expectations. In modern economies, contracts between individuals are typically incomplete[18–20] and parties often resort to renegotiation or, if this fails, to the courts. Partnerships and joint ventures are sustained by reciprocity[21,22], and fail when trust breaks down. In a sample of 92 joint ventures, Kogut et al.[22] found that only 32 were still functioning after 7 years. Marriage contracts too typically leave many aspects of the relationship unspecified[23–25] and, although customary law helps to align expectations, disputes and divorce are common[26,27]. Conversation and discourse analysis suggests that disagreements are routine in interpersonal interactions and that representing such conflict is a key part of human language[28–30]. Furthermore, individuals' perceptions about their partner's behaviour and their convictions about their own obligations are based on biased beliefs. An extensive literature in psychology has demonstrated that people are prone to self-serving biases in both causal attribution and judgements of fairness[31–35]. When interacting individuals are offered moral 'wiggle room' in experiments, they tend to choose the interpretation of the rules that serves their interest[36–39]. Such self-serving biases have been shown to prevent settlement in pre-trial bargaining between disputants[40–43], necessitating expensive third-party judicial intervention. Taken together, this evidence portrays a social world where partners often arrive at different conclusions regarding what has transpired, leading to disagreements that unravel reciprocal cooperation. While analytical work often assumes low error rates, simulation studies frequently assume substantial error rates. For example, two studies[2,44] incorporate error rates of 10% while a third[13] considers error rates of up to 50%. Laboratory experiments[16,17,45] investigating the effects of errors in the iterated prisoner's dilemma game assume error rates of 12.5–15%.

We introduce a strategy labelled arbitration tit-for-tat (ATFT) that uses third-party judgements to resolve disagreements about individual behaviour, and show that it can sustain cooperation even when error rates are high. Pairs of individuals, each characterized by a heritable strategy, are sampled from a large population and play an iterated prisoner's dilemma. In each interaction individuals can cooperate, producing benefit b to the partner at a cost c to themselves, or defect, generating no benefit or cost. Interactions continue with fixed probability w. With probability e, individuals whose strategy specifies that they should cooperate, instead defect but mistakenly believe that they have actually cooperated. Their partner perceives their actual behaviour.

We assume that behaviour takes place in a social group in which members share social norms and attend to the behaviour of others in a wide range of contexts such as marriage, child rearing,

[1]School of Human Evolution and Social Change, Arizona State University, Tempe, AZ, USA. [2]Institute of Human Origins, Arizona State University, Tempe, AZ, USA. ✉e-mail: Sarah.Mathew@asu.edu

sexual behaviour, sharing and exchange and participation in public goods provision. Group members make judgements about whether the behaviour of their peers conforms to the social norms that govern behaviour in these contexts. As a result, third parties have the knowledge necessary to arbitrate cases of pairwise cooperation. It is possible to show that such mutual monitoring behaviour can be evolutionarily stable[45], but it remains unknown whether these models explain real-world mutual monitoring of behaviour. Nonetheless, it is clear that in village-scale societies, people are intensely interested in, and aware of, each other's affairs and help to mediate interpersonal conflict and adjudicate disputes between well-known individuals. Amongst the Ju/'hoansi (!Kung) foragers of Botswana, conflicts over food sharing, a canonical form of reciprocity, are adjudicated by 'group talking' until a consensus is reached[46,47]. Amongst Turkana pastoralists, community discussion determines who has violated norms[48]. In Fijian villages, Arno[49] reports that talk around kava-drinking sessions is centred on what other people did, said and what their motives were, followed by discussions in which everyone present opines and interprets these actions; conflict between individuals in a culturally specified reciprocal relationship is adjudicated by third parties through a process of fact finding, application of norms, judgement and sanctioning. Various forms of informal third-party mediation that help to resolve conflicts between individuals in ongoing relationships are seen in small-scale agricultural and pastoral societies[50]. In a study of verbal conflicts within families, family members who were third parties to the conflict intervened in 38% of conflict episodes[51]. Even when formal litigation is an option, disputants who want to continue their association seek, and are satisfied by, informal non-coercive third-party mediation[52–54]. Here we take mutual monitoring as a given, and examine whether third-party adjudication can facilitate the evolution of cooperation even if norm violators are not being sanctioned.

The strategy ATFT is defined as follows: cooperate if your partner is in good standing[3] or you are in bad standing, otherwise defect. All individuals start in good standing, and remain so if they play by ATFT rules. When an individual perceives a defection that violates the ATFT rules, they call on a third party to judge whether the defection actually occurred. The arbitrator has a probability, a, of making an accurate judgement. If the arbitrator decides that a defection occurred that violates the ATFT rules, the player who defected falls into bad standing. If the arbitrator decides that a defection did not occur, or did occur but did not violate ATFT rules, the player who called for the arbitration procedure falls into bad standing. If an individual in bad standing cooperates and is not found to have invoked arbitration without cause, they return to good standing. If an individual in good standing defects with a partner in bad standing, the individual remains in good standing unless they are judged to have invoked arbitration without cause. Thus, ATFT individuals respect the social consensus and behave accordingly, even when they disagree with this consensus.

Arbitration creates a public signal about which players agree and, by conditioning standing on this signal, cooperation can persist even when players disagree about what actually happened. Suppose two ATFT players in good standing are interacting but one of them, labelled focal, mistakenly defects. The focal believes that she has cooperated, but their partner believes she defected. The partner asks a third-party arbitrator what occurred. In Table 1(top) the arbitrator correctly reports that the focal defected. Now both players agree that the focal is in bad standing during the next interaction, so the partner defects and the focal cooperates. Since both conformed to ATFT, both return to good standing. In Table 1(bottom) the arbitrator mistakenly agrees with the focal that she cooperated. As a result, during the next interaction the focal is in good standing and the partner is in bad standing, so the partner cooperates and the focal defects. Once again, both return to good standing. In effect, arbitration converts perception errors into execution errors by allowing individuals to condition their behaviour on their actual behaviour rather than on their perceptions, at least in expectation.

Table 1 | Sequence of possible moves by two ATFT players when the focal player makes an error (denoted by boldface) and the arbitrator is either accurate (top) or inaccurate (bottom)

Accurate arbitrator						
Arbitrator belief			**D**			
Focal standing	…	g	g	b	g	…
Focal behaviour	…	C	**D**	C	C	…
Partner behaviour	…	C	C	D	C	…
Partner standing	…	g	g	g	g	…
Inaccurate arbitrator						
Arbitrator belief			**C**			
Focal standing	…	g	g	g	g	…
Focal behaviour	…	C	**D**	D	C	…
Partner behaviour	…	C	C	C	C	…
Partner standing	…	g	g	b	g	…

Good standing is denoted by g, bad standing by b, cooperate by C and defect by D.

Results

A population in which ATFT is common can resist invasion by rare individuals using a range of different strategies, even if perception errors are common and arbitration is often inaccurate. Recent work on direct reciprocity has addressed the problem of evolutionary stability by defining a space of possible strategies and then testing those against all strategies possible in the set[2,15,55–57]. This approach is a substantial improvement over the ad hoc selection of strategies used in previous work. However, the strategy spaces that have been studied are relatively simple, precluding strategies such as ATFT which condition behaviour on the difference between intentions and behaviour or on external signals such as those generated by arbitration, and so we were not able to take this approach. Instead, we attack this problem in three ways. First, we show that if provision of arbitration is moderately accurate, ATFT is a strong subgame-perfect equilibrium: ATFT has higher expected fitness than any strategy that deviates from ATFT at every node in the game tree. This means that any strategy that follows ATFT, but occasionally does something different, has lower expected fitness and so cannot invade a population in which ATFT is common. Second, we derive the range of parameter values that allow ATFT to resist invasion by well-studied strategies. Third, we determine the memory-1 strategy best able to invade ATFT for a given set of parameters, and derive the conditions that make ATFT stable against that strategy. All three approaches suggest that ATFT has higher fitness than invasion strategies providing that arbitration is moderately accurate.

To prove that ATFT is subgame perfect, it is sufficient to show that, when paired with an ATFT player, single deviations from ATFT lead to lower expected pay-off at all equivalent nodes that two ATFT players will reach and at which deviation can occur[58]. There are two different classes of nodes: 'behaviour' nodes at which individuals choose whether to cooperate or defect and 'arbitration' nodes at which they choose whether or not to invoke arbitration. The standing of the focal and the partner are sufficient to determine the behaviour of ATFT at behaviour nodes. This means that there are three equivalence classes of behaviour nodes. Listing the focal's standing first, these are good–good (gg), good–bad (gb) and bad–good (bg). We omit bad–bad because ATFT choices at this node are identical to gg.

For each class, we determine conditions under which deviation from ATFT leads to lower expected pay-off. After individuals either cooperate or defect they can choose to call the arbitrator. If the focal chose cooperation, she does not know for sure that she cooperated and so can be at one of two nodes depending on whether an error occurred, leading to an information set, or sets if the partner also chose cooperation. To show that ATFT is subgame perfect, we show that it has a higher pay-off at each of these information sets than a 'deviant' strategy that deviates from ATFT—that is, it calls the arbitrator when ATFT does not and does not call the arbitrator when ATFT does.

To determine when ATFT has higher fitness than strategies that deviate from ATFT, we derive analytical expressions for the expected fitness of ATFT playing another ATFT individual at each type of node. These expressions are then used to derive analytical expressions for the expected fitness of individuals who deviate from ATFT at either behaviour or arbitration nodes, assuming that the deviants obey ATFT except for their deviation. While the expected fitness of an ATFT individual paired with another ATFT individual is almost independent of the level of arbitration accuracy (Supplementary Fig. 1), the expected fitness of deviating strategies when interacting with ATFT is strongly affected by arbitration accuracy.

Here we present the results for the two types of deviation: individuals who defect when both players are in good standing (DD deviants) and individuals who do not call the arbitrator and subsequently cooperate when their partner mistakenly defects (CD deviants). We show (Supplementary Note 2) that, if these deviations lead to lower expected pay-off than that obtained by ATFT, then so do all other deviations and ATFT is a subgame-perfect equilibrium strategy for the range of parameters considered.

A DD deviant who defects when both players are in good standing has lower fitness than one who cooperates for the combinations of arbitration accuracy and error rate shown in Fig. 1. When perception error rates are low, ATFT has higher expected fitness providing arbitration is slightly better than random. As errors become more common the minimum arbitration accuracy increases but, even when perception error rates are as high as 50%, ATFT has higher expected fitness when arbitration is correct only 90% of the time.

Figure 2 plots the minimum arbitration accuracy necessary for ATFT to have higher expected fitness than a CD deviant who does not call the arbitrator when her partner mistakenly defects at a gg–cc node. Because the arbitrator is not called, the partner remains in good standing and the focal cooperates during the next interaction. When ATFT perceives a defection, it calls the arbitrator and then retaliates against the defector if the arbitrator does not err. However, calling the arbitrator comes at a cost. When the arbitrator errs, the ATFT individual calling the arbitrator falls into bad standing and loses the benefit of cooperation during the next interaction. As a result, when errors are rare, arbitration must be moderately accurate for ATFT to have higher expected fitness than individuals who ignore the arbitrator and simply cooperate but, as errors become more common, the arbitration accuracy required declines.

The key insight from our analysis is that arbitration need not be highly accurate to stabilize ATFT against invasion by any strategy that deviates from ATFT. When errors are common, arbitration accuracy must be about 0.8 for ATFT to resist invasion by a strategy that defects rather than cooperates (for the parameters studied here). Similarly, when errors are rare, arbitration accuracy must be approximately $\frac{b}{b+c}$, which is 0.75 when $b/c=3$, if ATFT is to resist invasion by a strategy that does not call the arbitrator after defections and subsequently cooperates. Note that when $a=0.5$ arbitration is random: it gives the focal no information about the behaviour of her partner; thus $a=0.75$ is midway between completely uninformative and completely accurate arbitration. For intermediate error rates, the minimum accuracy necessary to resist invasion by strategies that deviate from ATFT is lower but still >0.5. This means

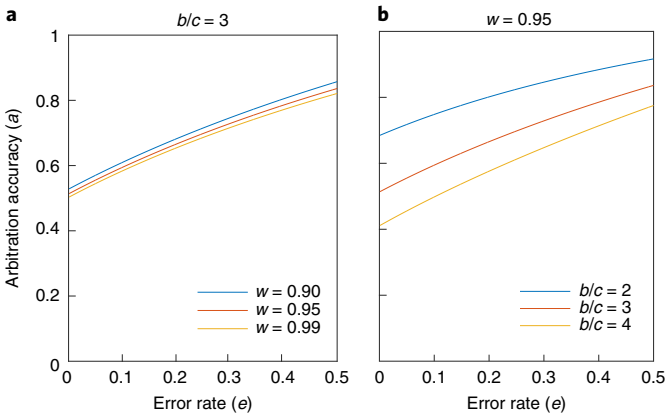

Fig. 1 | The minimum arbitration accuracy necessary for ATFT to have higher fitness than a deviant who defects at the gg node, as a function of error rate. Shown are three different continuation probabilities (left) and benefit/cost ratios (right). In all cases, $c=1$.

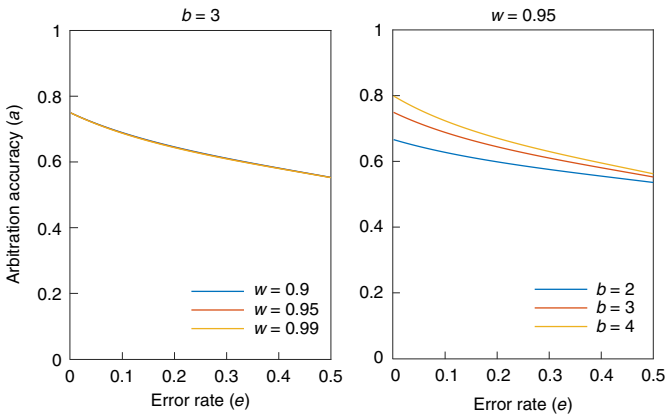

Fig. 2 | Minimum arbitration accuracy necessary for ATFT to have higher expected fitness than an individual who does not call the arbitrator after the partner mistakenly defects and who cooperates on the next interaction because the partner remains in good standing. Shown are three different continuation probabilities (left) and benefit/cost ratios (right). In all cases, $c=1$.

that using a random arbitration convention, such as flipping a coin, will not work: judgements must be correlated with actual behaviour. Also, note that these results are not simply a consequence of the folk theorem: the folk theorem holds that strategies have pay-off greater than or equal to alternative strategies, and so such strategies are subject to indirect invasion[4]. Here, there is no indirect invasion by modified strategies because we require ATFT to have strictly higher expected pay-off than alternatives.

Next we consider several plausible strategies that are not small modifications of ATFT: ALLD, Cheater, ALLC, Tolerant, WSLS, GTFT and tit-for-tat with arbitration (TFTA). We derive analytical expressions for the expected fitness of these strategies when paired with an ATFT individual and, using these expressions, we examine the range of conditions under which these strategies can invade a population comprised of ATFT individuals. ALLD individuals defect on every move, and do not consult the arbitrator when their partner defects. The conditions for ATFT to resist invasion by ALLD are qualitatively similar to, but less restrictive than, those for ATFT to have higher expected fitness than a strategy that

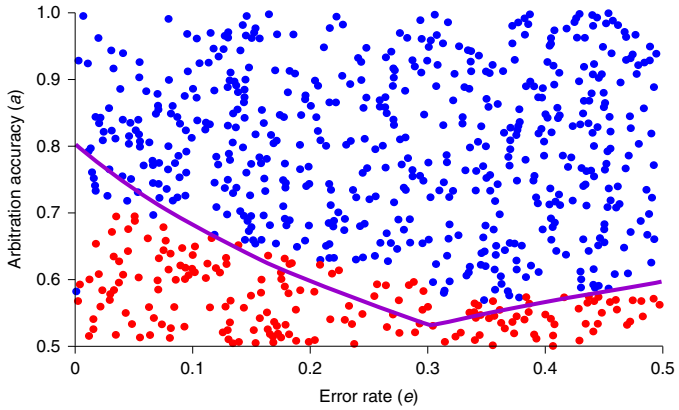

Fig. 3 | ATFT resists invasion by the best memory-1 strategy. The blue dot plot shows combinations of *e* and *a* at which the best memory-1 strategy has a lower expected fitness when interacting with ATFT than does ATFT when interacting with ATFT. The red dots are combinations at which ATFT has lower expected fitness. Above the purple line, ATFT can resist invasion by ALLC and ALLD while below it cannot. $b=3$, $w=0.95$.

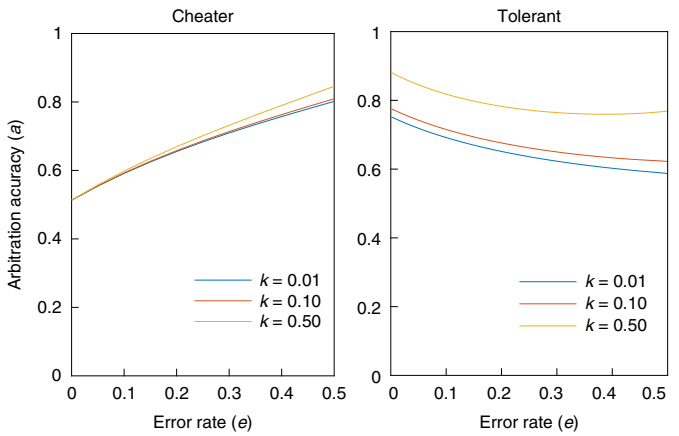

Fig. 4 | The minimum arbitration accuracy necessary to resist invasion by Cheater or Tolerant as a function of error rate for three levels of arbitration cost. Assuming the arbitration cost is not a substantial fraction of the cost of cooperation, the results do not qualitatively differ from those without arbitration costs. In all cases, $b=3$, $c=1$ and $w=0.95$.

deviates from ATFT at the gg node (Fig. 1). The Cheater strategy defects with probability d in circumstances in which ATFT specifies that it should cooperate; otherwise it conforms to ATFT. When d is small, Cheater defects only rarely. As d increases, Cheater defects more often until, when $d=1$ it is similar to ALLD except that it consults the arbitrator when its partner defects. Again, the minimum arbitration accuracy that allows ATFT to resist invasion by Cheater when d is small is very similar to what is needed for ATFT to have higher expected fitness than a strategy that deviates from ATFT at the gg node (Fig. 1). Increasing d reduces the expected fitness of Cheater when interacting with ATFT. ALLC individuals cooperate on every move, do not call the arbitrator when their partner defects and ignore the decisions of the arbitrator called by their partner. The conditions for ATFT to resist invasion by ALLC are qualitatively similar to, and less restrictive than, the conditions for ATFT to resist invasion by the strategy that does not call the arbitrator (Fig. 2). Tolerant individuals do not consult the arbitrator and cooperate in the next interaction, even when they perceive that their partner defected in violation of ATFT rules. Unlike ALLC, Tolerant individuals pay attention to the decision of the arbitrator called by their partner. The conditions for ATFT to resist invasion by Tolerant are similar to those to resist invasion by ALLC. WSLS cooperates when one player cooperates and the other defects. ATFT can resist invasion by WSLS providing arbitration accuracy is >0.5. GTFT cooperates after a defection with probability g. Increasing g always increases the average pay-off of GTFT interacting with GTFT but, if g is too large, GTFT can be invaded by defecting strategies. For the parameters used here, the maximum value of g varies from about 0.6 when errors are rare to <0.4 when they are common. At high values of g, the conditions for ATFT to resist invasion by GTFT are similar to, but less restrictive than, those that allow ATFT to resist invasion by ALLC. As g increases, the conditions become even less restrictive. TFTA plays tit-for-tat, except that when it perceives that its partner has defected it calls the arbitrator and accepts the arbitrator's decision about its partner's behaviour. If the arbitrator says that its partner cooperated, TFTA cooperates on the next move; if the arbitrator says its partner defected, TFTA defects on the next move. Therefore, like ATFT, an individual playing this strategy conditions her behaviour on information provided by the arbitrator about her partner. However, unlike ATFT, this strategy does not modify its behaviour based on its own standing or its partner's standing. ATFT can resist invasion by TFTA for reasonable arbitration accuracy levels providing the perception error rate is <0.5. Note that, within this set of strategies, there is no indirect invasion or cycling. Once common, ATFT is stable against all of the other strategies.

Finally, we consider invasion by the best memory-1 strategy for any given set of parameter values. We randomly chose an error rate (e) uniformly distributed between 0 and 0.5 and an arbitration accuracy (a) uniformly distributed between 0.5 and 1, and used the Matlab optimization function fminsearch to find the memory-1 strategy with the highest expected pay-off when interacting with ATFT. When the fitness of the best memory-1 strategy was less than the expected fitness of ATFT interacting with ATFT, we considered ATFT to be an evolutionarily stable strategy (ESS) against that memory-1 strategy. The results (Fig. 3) suggest that, for error rates and arbitration accuracies that allow ATFT to resist invasion by ALLC and ALLD, it can also resist invasion by the best memory-1 strategy.

So far we have assumed that appealing to arbitration has no cost because people monitor the affairs of community members for other reasons. While this could be a fair approximation for life in a village, in some settings arbitrators may require compensation for their services. We modelled arbitration costs by assuming that each time an individual calls the arbitrator she experiences a pay-off reduction k. Figure 4 shows the effect of arbitration cost on the minimum arbitration accuracy necessary for ATFT to resist invasion by Cheater, the strategy that intentionally defects with low probability, and by Tolerant, the strategy that does not call the arbitrator after errors but then cooperates. Arbitration cost has little effect on the ability to resist invasion by Cheater, but it does affect the ability of ATFT to resist invasion by Tolerant depending on the size of arbitration cost relative to the cost of cooperation. Arbitration cost has little effect when $k=0.1c$, but requires arbitration accuracy of close to 90% when $k=0.5c$.

In real-world interactions, arbitration may not only be inaccurate but may also be biased. We analyse whether ATFT can persist if arbitration is biased in favour of certain individuals and against others. We assume that there are two types of individual—central and peripheral. When a central and a peripheral individual interact, arbitrators are more likely to judge the former's mistaken defection to be cooperative rather than the mistaken defection of a peripheral individual. This means that a peripheral individual is more likely to fall into bad standing after an error and so will have lower pay-offs. Nonetheless, under a wide range of conditions, peripheral individuals are better off cooperating when paired with a central

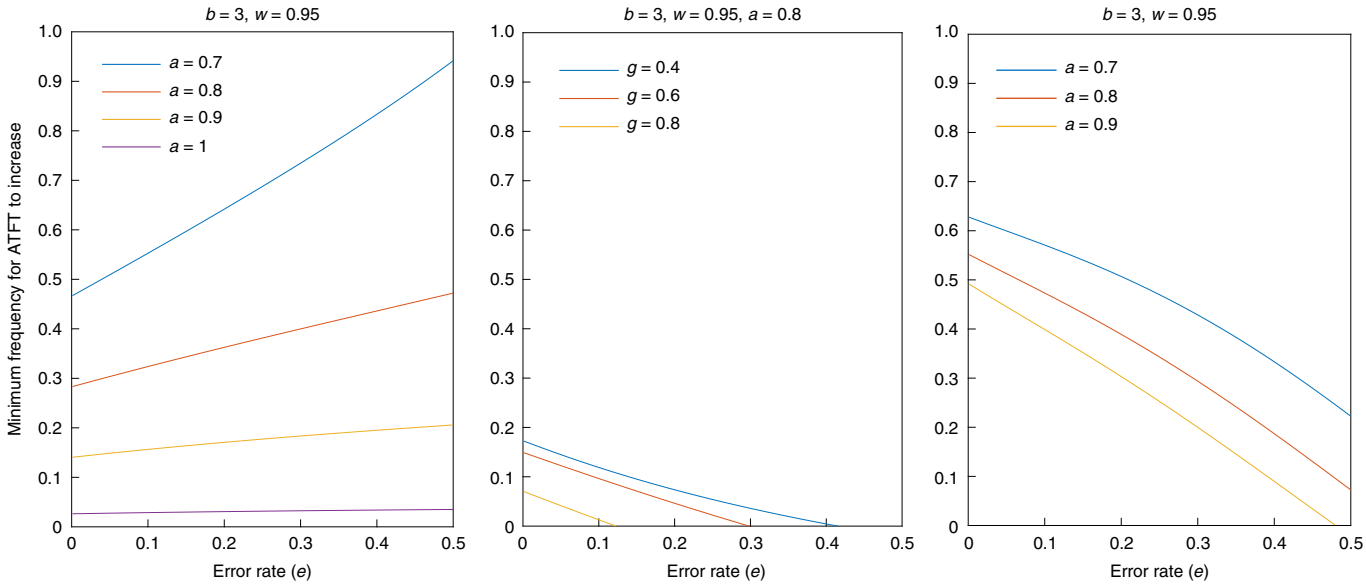

Fig. 5 | Invasion criteria for ATFT. The minimum frequencies of ATFT necessary for it to increase in competition with ALLD (left), GTFT (middle) or WSLS (right), as a function of error rate. Against ALLD, the minimum frequencies are very sensitive to arbitration accuracy. For the reciprocating strategies GTFT and WSLS, the minimum frequencies are very sensitive to error rate. In all cases, $b = 3$, $c = 1$ and $w = 0.95$.

individual than when they are not interacting with that individual. Additionally, peripheral individuals are better off choosing to cooperate with a central individual rather than with another peripheral individual, providing cooperation with a central individual provides modestly higher pay-offs.

ATFT increases in frequency when in competition with ALLD, GTFT and WSLS once it reaches a minimum frequency (Fig. 5). The minimum frequency needed to invade ALLD is very sensitive to arbitration accuracy: high accuracy makes invasion easy while low accuracy makes it difficult. Increasing the benefit/cost ratio lowers the minimum frequency. The minimum frequencies needed to invade GTFT or WSLS are very sensitive to the error rate. At low error rates, ATFT must achieve a modest frequency before it increases; at higher error rates, WSLS and especially GTFT are more easily invaded. For certain parameter combinations, ATFT invades populations comprised of either GTFT or WSLS even when ATFT is rare. In all cases, once ATFT invades it evolves to high frequency.

Discussion

We have shown that, by relying on a public, third-party opinion to align players' beliefs, ATFT can re-establish cooperation even when errors arise frequently. The arbitration can be imperfect but not random. ATFT can resist invasion by strategies that occasionally or always defect, by strategies that call the arbitrator dishonestly when no error has occurred, and by second-order free-riding strategies that cooperate or fail to call the arbitrator when a defection has occurred. Higher error rates increase the range of conditions under which ATFT can resist invasion by such overly cooperative strategies. These results suggest that ATFT is a plausible evolutionary strategy that warrants further analysis.

A limitation of our analysis is that we have not been able to show that ATFT can resist all possible invaders in a space of strategies that includes ATFT. Current methods to perform such analyses could not be extended to our case because ATFT is a state-dependent strategy that conditions behaviour on what was intended and what occurred. Instead, we derived the conditions under which ATFT is a strong subgame-perfect equilibrium—that is, ATFT has higher fitness than any strategy that deviates from ATFT at any node along the equilibrium path. Second, we simulated the conditions that allow ATFT to resist the best memory-1 invader. Third, we analytically derived the conditions under which ATFT has higher expected fitness than plausible and well-known strategies. All three approaches suggest that ATFT is evolutionarily stable under a range of plausible conditions. However, we cannot exclude the possibility that ATFT can be invaded by strategies that are more complex than memory-1 strategies and that use arbitration but are substantially different from ATFT.

The use of third-party arbitration to resolve perception errors may shed light on why reciprocity plays a more important role in humans than in other social vertebrates. Division of labour and exchange, food sharing and mutual aid are widespread in human societies, and reciprocity is one key mechanism for maintaining such cooperation[59–65]. In other vertebrates the evidence for reciprocity among conspecifics is sparse and contentious[66–73]. The dearth of reciprocal cooperation in nature is puzzling, since many species live in social groups in which individuals recognize other group members, interact repeatedly over time and seem to have the cognitive ability to adjust current behaviour contingent on another's past action, exactly the conditions that should favour reciprocity[1]. It is possible that reciprocity is rare in these species because perception errors are common. When errors arise frequently, strategies such as WSLS and GTFT do not perform well (Supplementary Note 7). Third-party arbitration is rare outside of our species, and where it exists, it takes the form of interventions in agonistic interactions, not of informing potential cooperators[74–76]. This may preclude the evolution of strategies such as ATFT.

The benefits of using third-party arbitrators may also explain why reciprocity in humans tends to be regulated by shared social norms supported by third-party intervention[77]. In small-scale societies, disputes between friends, couples or neighbours are resolved by raising the issue with other friends, family members or elders in the community, who discuss the matter, offer their opinion, pass judgements and help mediate a consensus[47,77]. Formal courts also regularly settle disputes between domestic partners, employer and employee and co-workers. Our results can explain why pairwise relationships built on a long history of repeated interactions cannot be maintained by the threats of defection alone, but require the scaffolding of third-party judgements and shared social norms.

Methods

We used analytical and simulation methods to evaluate the conditions under which ATFT can persist. The calculations we did can be divided into the following seven parts: (1) we derived the expected fitness of ATFT playing against itself. (2) We determined the conditions under which ATFT is a strong subgame-perfect equilibrium. (3) We examined the conditions under which ATFT can resist invasion by plausible invading strategies, namely ALLD, ALLC, Cheater, Tolerant, GTFT, WSLS and TFTA. (4) We numerically calculated the combinations of error rate and arbitration accuracy necessary for ATFT to resist the best memory-1 invader. (5) We examined the effects of making arbitration costly. (6) We computed the conditions under which individuals who suffer arbitration bias will nonetheless rely on arbitration. (7) We computed the initial frequency that allows ATFT to increase when in competition with ALLD, GTFT and WSLS. Complete calculations are provided in the Supplementary Information.

Expected fitness of ATFT playing against itself. We derived the expected fitness of ATFT playing against itself (Supplementary Note 1), which was used in subsequent analyses to assess the conditions under which ATFT can be evolutionarily stable. The behaviour of ATFT is determined by its assessment of its own standing and that of its partner. For example, if both are in good standing, both intend to cooperate. Each player errs with probability e and, when this occurs, the other individual calls the arbitrator who is correct with probability a. Each combination of error and arbitration event determines the players' standing on the next interaction, which occurs with probability w. The possible events and pay-offs are given in Supplementary Table 1. Using these, we derived an expression for the expected fitness of an ATFT individual when both she and her partner are in good standing, which is a linear equation in the expected fitnesses of an ATFT individual in each combination of her own and partner's standing. Supplementary Tables 2 and 3 give the events and pay-offs when the focal is in good standing and the partner is in bad standing, and when the focal is in bad standing and the partner is in good standing. Using these tables we get expressions for the expected fitness of ATFT for each combination of standings. This yields three linear equations in three unknowns (equations (1)–(3) in the Supplementary Information). There are only three equations because the fitness of ATFT when both players are in bad standing is the same as when both are in good standing. These three equations were solved using Mathematica to generate an expression for the expected fitness at the beginning of an interaction (equation (4) in the Supplementary Information). These expressions were then converted to Matlab expressions using the Mathematica function ToMatlab. Plots were generated in Matlab.

Conditions under which ATFT is a strong subgame-perfect equilibrium. To prove that ATFT is subgame perfect, we showed that a strategy that makes a single deviation from ATFT when paired with an ATFT player has lower expected pay-off at all equivalent nodes at which such deviations can occur (Supplementary Note 2). The standing of the focal and partner are sufficient to determine the behaviour of ATFT, so there are three equivalence classes of nodes at which individuals choose whether to cooperate or defect. Listing the focal's standing first, these are gg, gb and bg. We omit bad–bad because ATFT choices at this node are identical to those made at gg. If an individual chooses to cooperate, an error may occur causing that individual to defect. Behaviour at each of these nodes leads to a decision on whether to call the arbitrator. If the focal chose cooperation, she can then be at one of two nodes depending on whether an error occurred, leading to an information set, or sets if the partner also chose cooperation (see the game trees shown in Supplementary Figs. 6–8). To show that ATFT is subgame perfect, we determine the range of parameter values at which ATFT has a higher pay-off at each of these information sets than a 'deviant' strategy, which does the opposite to ATFT—calls the arbitrator when ATFT does not and does not call the arbitrator when ATFT does. The pay-offs and events are listed in Supplementary Tables 4–6. These expected fitnesses can be calculated using the stage pay-offs shown in the game trees, and the expected fitnesses of ATFT at the four decision nodes that were described in the first paragraph of Methods. Deviation may also occur at arbitration nodes. Sometimes the focal is not certain about which node she reached and so expected pay-offs have to be computed taking this into account. The events and stage pay-offs are given in Supplementary Tables 7–11, and the resulting expected fitness expressions are shown in Supplementary Information. These are complex expressions, and so the Mathematica output was converted to Matlab code using ToMatlab. The range of parameter values at which ATFT had higher fitness than the deviant was calculated numerically in Matlab using the function fzero, and is displayed in Figs. 1 and 2 and Supplementary Figs. 2–5.

Expected fitness of plausible invading strategies when paired with ATFT. We considered the following invading strategies:

1. ALLD individuals always defect and do not call the arbitrator when their partner defects.
2. Cheaters play ATFT except that in each interaction with independent probability d they intentionally defect. By varying d we can examine the effect of the rate of defection.
3. ALLC individuals always cooperate and never appeal to the arbitrator.
4. Tolerant individuals who perceive an error by their partner do not appeal to the arbitrator but, when the partner appeals to the arbitrator, they obey the rules of ATFT.
5. WSLS individuals cooperate when they perceive that during the last interaction both individuals cooperated or defected; otherwise they defect.
6. GTFT individuals cooperate when their partner cooperated during the previous interaction and with probability g when she defected; otherwise they defect.
7. TFTA individuals play tit-for-tat, meaning that they do whatever they think their partner did on the previous round. When their partner defects, they consult the arbitrator and rely on the arbitrator's assessment of their partner's behaviour.

In each case we computed the range of parameter values that cause the strategy to have lower expected fitness when paired with ATFT than ATFT has with itself (Supplementary Note 3). The calculations for ALLD, ALLC, Cheater and Tolerant are similar to those given in the first paragraph of Methods. ALLD and ALLC do not condition their behaviour on the behaviour of the other player; Cheater and Tolerant condition their behaviour using the same standing rules as ATFT, and so in each case there are four equivalence classes of behaviour nodes. We derived an expression for the expected fitness of the invading strategy at each type of node in terms of the fitnesses at other nodes, and then solved the resulting system of four simultaneous linear equations using Mathematica (Supplementary Tables 12–20). The minimum values of a that allow ATFT to resist invasion were calculated using these fitnesses and the Matlab function fzero, and are plotted in Supplementary Figs. 9–12.

The computation of expected fitness for WSLS, GTFT and TFTA is more complex. In each case, the invading strategy makes current behaviour contingent on past choices but uses a different rule than ATFT. For example, WSLS cooperates if both individuals cooperated or both defected on the previous interaction; otherwise it defects. This means that the fitness of a WSLS individual depends on its own intentions and that of its ATFT partner. We therefore calculated the expected fitness of WSLS at each combination of the WSLS player's intention to cooperate or defect, and the WSLS player's standing as viewed by the ATFT player. This leads to five simultaneous linear equations (Supplementary Note 3 and Supplementary Tables 21–26). Solving these equations using Mathematica yields an expression for the expected fitness of WSLS. With these fitness expressions and using the Matlab function fzero, we numerically calculated the minimum values of a for ATFT to resist invasion by WSLS.

The strategy GTFT cooperates if its partner cooperated on the previous round. If the partner defected, GTFT cooperates with probability g, and defects with probability $1-g$. As usual, ATFT's behaviour depends on it's assessment of the standings of the two players. As with WSLS, the expected fitness of GTFT depends on both its intention to either cooperate or defect, and the standings of the two individuals. For each combination, the behaviour of each individual determines the stage pay-offs, the intention of the GTFT individual and the standing of both in the next time period should it occur (Supplementary Tables 27–32). This yielded five simultaneous equations that can be solved for the expected fitness of GTFT at the beginning of an interaction. The expected fitness of GTFT depends on the parameter g, the probability that GTFT cooperates after a defection. The value of g that we used when assessing the conditions under which ATFT can resist invasion by GTFT is its maximum value at which GTFT can resist invasion by ALLD. We computed the fitness of ALLD interacting with GTFT and used the Matlab function fzero.m to determine this value of g.

The strategy TFTA cooperates if its partner cooperated during the previous interaction. If its partner defected, TFTA consults the arbitrator. If the arbitrator rules that the partner cooperated, TFTA cooperates, otherwise it defects. This complicates computation of expected fitnesses because the arbitrator's decision may affect the TFTA individual and her ATFT partner differently. Supplementary Tables 33–38 give the events and stage pay-offs for each combination of TFTA's intent and ATFT's belief about standing. These generated a system of linear equations that were solved using Mathematica and were used to generate Supplementary Fig. 16 in Mathematica.

Conditions required for ATFT to resist invasion by the best memory-1 strategy. We investigated which combinations of error rate and arbitration accuracy would allow ATFT to resist invasion by the best memory-1 strategy (Supplementary Note 4). A memory-1 strategy is defined by five probabilities: the probability of cooperating on the first round, and the probabilities of cooperating after each of the four possible behaviours by both interacting individuals on the previous round—DD, CD, DC and CC. We randomly chose an error rate e, uniformly distributed between 0 and 0.5, and an arbitration accuracy a, uniformly distributed between 0.5 and 1, and used the Matlab optimization function fminsearch to find the memory-1 strategy with the highest expected pay-off when interacting with ATFT. We calculated the expected fitness of a memory-1 strategy interacting with ATFT, by simulating 100,000 interactions and averaging the pay-offs. We repeated the optimization procedure five times with different initial values and took the largest of the five values to be the best memory-1 strategy. When the fitness of the best memory-1 strategy was less than the expected fitness of ATFT interacting with ATFT, we considered ATFT to be an ESS against that memory-1 strategy. The results are plotted with Excel in Fig. 4.

Eight Matlab scripts were used for this analysis, and are available at https://osf.io/weu4b/?view_only=7fb48e283424447c930d0455aaa36912:

1. BestMemOneRndPlot.m: base script. Sets up parameter values to be simulated, calls bestdiffmin and outputs results to a text file (Supplementary Fig. 3).
2. MinAEssVsAllcAlld.m: computes the minimum value of a required for ATFT to resist invasion by ALLC and ALLD, given w, e, b, using fzero and DiffPayAtftAllc.m and DiffPayAtftAlld.m.
3. DiffpayAtftAllc.m: calculates the difference in expected pay-offs between ATFT playing against itself and ALLC playing against ATFT.
4. DiffPayAtftAlld.m: calculates the difference in expected pay-offs between ATFT playing against itself and ALLD playing against ATFT.
5. bestdiffFmin.m: uses MatLab function fminsearch to find the five MemOne probabilities that minimize the difference in pay-off between MemOne and ATFT. fminsearch uses the simplex algorithm to find a local minimum in diftrans.m. Convergence requires that the minimum value changes less than Tolfun and that all elements of the behaviour matrix change less than TolX. Values plotted in Fig. 3 are based on Tolfun = TolX = 0.01. Smaller values greatly increased the computation time and led to only very small changes in pay-off of the best memory-1 strategy.
6. diftrans.m: called by fminsearch. Translates the 1×5 vector supplied by fminsearch into initial behaviour probability and the 2×2 conditional behaviour matrix required by EpayAtftVsMemOne.m. Calls EpayAtftVsMemOne and subtracts ATFT's pay-off from that of the memory-1 strategy given by MemOne, $x(1) = \text{InitC} = P(c \mid \text{first move})$, $x(2) = \text{Mem1Mat}(1,1) = P(c \mid dd)$, $x(3) = \text{Mem1Mat}(1,2) = P(c \mid dc)$, $x(4) = \text{Mem1Mat}(2,1) = P(c \mid cd)$, $x(5) = \text{Mem1Mat}(2,2) = P(c \mid cc)$.
7. EpayAtftVsMemOne.m: calculates the per-period expected pay-offs for an ATFT player interacting with a memory-1 player in a prisoner's dilemma with an arbitrator.
8. PayAtftVsMemOne.m. Calculates the pay-off for ATFT against a given memory-1 player during one interaction with t periods.

When arbitration is costly. We investigated the effect of adding a cost to arbitration by assessing how arbitration costs affect the conditions for ATFT to resist invasion by Cheater and Tolerant (Supplementary Note 5). Each time the arbitrator is called, the individual calling the arbitrator pays a cost k. We calculate the expected fitnesses of ATFT playing against ATFT, Cheater playing against ATFT and Tolerant playing against ATFT. ATFT individuals who intend to cooperate defect with probability e. Because Cheaters introduce intentional defections, their probability of defecting is $e' = e + d$. We calculate the expected fitness of ATFT by setting $d = 0$. As in the model without arbitration costs, a player who experiences a defection invokes the arbitrator. The arbitrator decides whether the partner of the player who invokes the arbitration has defected, accurately with probability a and inaccurately with probability $1 - a$. When an individual calls the arbitrator she pays a cost k to those doing the arbitration. As usual, the behavioural decisions of both ATFT and Cheater depend on the standing of the two individuals who are interacting. Supplementary Tables 39–41 give the events and pay-offs for two individuals when both are in good standing, when the focal is in bad standing and when the partner is in bad standing. These yield three linear equations, which are solved using Mathematica to find the expected fitness of individuals interacting with ATFT when there are arbitration costs. These expressions can be used to compute the fitness of ATFT by setting $d = 0$, and of Cheater by setting d to a small value.

A Tolerant individual does not call the arbitrator when she believes that her partner has defected, but instead ignores the defection and acts as if her partner were still in good standing. The ATFT player who experiences a defection invokes the arbitrator and experiences a cost k, but the Tolerant focal does not. Both players condition their behaviour on the standing of both players, but it is not possible for both players to be in bad standing simultaneously. ATFT falls into bad standing only if it invokes the arbitrator when its Tolerant partner mistakenly defects and the arbitrator inaccurately rules that the partner did cooperate. When this occurs, the Tolerant partner goes into good standing. Supplementary Tables 42–45 give the events and pay-offs for each feasible combination of standing. These led to three simultaneous linear equations, which were solved using Mathematica. The minimum value of a that allowed ATFT to resist invasion was found using the Matlab function fzero for a range of arbitration costs, as shown in Fig. 4 and Supplementary Figs. 17 and 18.

Biased arbitration. We considered whether cooperation can occur if the arbitration process systematically favours one class of people (central individuals) over members of another class (peripheral individuals) (Supplementary Note 6). We assume that, when central individuals play with peripheral individuals, the arbitration process favours central individuals so that when they defect, the arbitrator is more likely to say that they cooperated. We also assume that a central individual can confer benefits to a peripheral individual that a peripheral person cannot, so that when a peripheral and central interact, the peripheral member gets benefit $b + d$ at cost c, and the central gets benefit b at cost c. We assumed that the peripheral individual can choose whether or not to enter into a cooperative arrangement with another individual.

We considered two possible situations. The first was that peripheral individuals can choose between a central and a peripheral partner. Then ATFT is stable providing the pay-off to a peripheral interacting with a central is greater than that to a peripheral interacting with a peripheral. The second was that peripheral individuals have to choose between a central partner or no partner. In this case, ATFT is stable if the pay-off of a peripheral individual interacting with a central individual is greater than zero. Again the behaviour of both types depends on their standings. Supplementary Tables 46–48 give the events and actions for each combination of standings. Using these tables we derived a system of three simultaneous equations (Supplementary Note 6) and solved them using Mathematica to yield an expression for the expected fitness of a peripheral individual. We used Mathematica to generate Supplementary Figs. 19 and 20.

Conditions for ATFT to invade ALLD, GTFT and WSLS. We calcuated the expected fitnesses of a focal ATFT individual when it plays against ALLD, GTFT and WSLS, and examined the frequency of ATFT at which ATFT has the same fitness as these strategies in populations containing ATFT and one of these strategies with players being paired at random (Supplementary Note 7). Since ALLD never cooperates in interactions with other ALLD individuals, the expected fitness of ALLD when playing against itself is zero. We calculated the expected fitness of an ATFT focal player when paired with an ALLD player using the events and pay-offs shown in Supplementary Tables 49–51. We used this, along with the pay-off of ATFT playing against itself, to compute the average fitness of ATFT as a function of the frequency of ATFT in the population. With the Matlab function fzero we determined the frequency at which this average fitness is zero. We used a similar procedure to calculate the fitness of ATFT when paired with WSLS and GTFT. The events and behaviours are detailed in Supplementary Tables 52–58 and the resulting equations are displayed in Supplementary Note 7. These calculations were used to produce Fig. 5.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

There is no empirical data associated with this paper.

Code availability

Mathematica and Matlab scripts used to solve the fitness equations, perform Monte Carlo simulations and create the plots are publicly available at https://osf.io/weu4b/?view_only=7fb48e283424447c930d0455aaa36912.

Received: 24 May 2019; Accepted: 27 October 2020;
Published online: 04 January 2021

References

1. Trivers, R. L. The evolution of reciprocal altruism. *Q. Rev. Biol.* **46**, 35–57 (1971).
2. van Veelen, M., García, J., Rand, D. G. & Nowak, M. A. Direct reciprocity in structured populations. *Proc. Natl Acad. Sci. USA* **109**, 9929–9934 (2012).
3. Sugden, R. *The Economics of Rights, Co-operation and Welfare* (B. Blackwell, 1986).
4. Boyd, R. Mistakes allow evolutionary stability in the repeated prisoners-dilemma game. *J. Theor. Biol.* **136**, 47–56 (1989).
5. Boerlijst, M. C., Nowak, M. A. & Sigmund, K. The logic of contrition. *J. Theor. Biol.* **185**, 281–293 (1997).
6. Nowak, M. & Sigmund, K. The evolution of stochastic strategies in the prisoner's dilemma. *Acta Appl. Math.* **20**, 247–265 (1990).
7. Nowak, M. & Sigmund, K. Evolution of indirect reciprocity. *Nature* **437**, 1291–1298 (2005).
8. Nowak, M. A. & Sigmund, K. A strategy of win–stay, lose–shift that outperforms tit-for-tat in prisoner's dilemma. *Nature* **364**, 56–58 (1993).
9. Posch, M. Win–stay, lose–shift strategies for repeated games–memory length, aspiration levels and noise. *J. Theor. Biol.* **198**, 183–195 (1999).
10. Imhof, L. A., Fudenberg, D. & Nowak, M. A. Tit-for-tat or win–stay, lose–shift? *J. Theor. Biol.* **247**, 574–580 (2007).
11. Molander, P. The optimal level of generosity in a selfish, uncertain environment. *J. Conflict Resolut.* **29**, 611–618 (1985).
12. Nowak, M. & Sigmund, K. Tit for tat in heterogeneous populations. *Nature* **355**, 250–253 (1992).
13. Zagorsky, B. M., Reiter, J. G., Chatterjee, K. & Nowak, M. A. Forgiver triumphs in alternating prisoner's dilemma. *PLoS ONE* **8**, e80814 (2013).
14. Boyd, R. & Lorberbaum, J. P. No pure strategy is evolutionarily stable in the repeated prisoners-dilemma game. *Nature* **327**, 58–59 (1987).
15. Hilbe, C., Chatterjee, K. & Nowak, M. A. Partners and rivals in direct reciprocity. *Nat. Hum. Behav.* **2**, 469–477 (2018).
16. Van Lange, P. A., Ouwerkerk, J. W. & Tazelaar, M. J. How to overcome the detrimental effects of noise in social interaction. *J. Pers. Soc. Psychol.* **82**, 768–780 (2002).

17. Fudenberg, D., Rand, D. G. & Dreber, A. Slow to anger and fast to forgive: cooperation in an uncertain world. *Am. Econ. Rev.* **102**, 720–749 (2012).
18. Williamson, O. *The Economic Institutions of Capitalism* (Macmillan, 1985).
19. Hart, O. & Moore, J. Incomplete contracts and renegotiation. *Econometrica* **56**, 755–785 (1988).
20. Hart, O. & Moore, J. Foundations of incomplete contracts. *Rev. Econ. Stud.* **66**, 115–138 (1999).
21. Axelrod, R. M. *The Evolution of Cooperation* (Basic Books, 1984).
22. Kogut, B., Kogut & Bruce The stability of joint ventures: reciprocity and competitive rivalry. *J. Ind. Econ.* **38**, 183–198 (1989).
23. Weitzman, L. J. Legal regulation of marriage: tradition and change: a proposal for individual contracts and contracts in lieu of marriage. *Calif. Law Rev.* **62**, 1169 (1974).
24. Goodale, J. C. Marriage contracts among the Tiwi. *Ethnology* **1**, 452–466 (1962).
25. Dnes, A. W. & Rowthorn, B. *The Law and Economics of Marriage and Divorce* (Cambridge Univ. Press, 2002).
26. Cahn, D. D. *Conflict in Intimate Relationships* (Guilford Press, 1992).
27. Betzig, L. Causes of conjugal dissolution: a cross-cultural study. *Curr. Anthropol.* **30**, 654–676 (1989).
28. Briggs, C. L. *Disorderly Discourse: Narrative, Conflict and Inequality* (Oxford Univ. Press, 1996).
29. Duranti, A. in *Disentangling: Conflict Discourse in Pacific Societies* (eds Watson-Gegeo, K. A. & While, G. M.) 459–489 (Stanford Univ. Press, 1990).
30. Brenneis, D. Telling troubles: narrative, conflict and experience. *Anthropol. Linguist.* **30**, 279–291 (1988).
31. Haidt, J. *The Righteous Mind: Why Good People are Divided by Politics and Religion* (Pantheon, 2013).
32. Shepperd, J., Malone, W. & Sweeny, K. Exploring causes of the self-serving bias. *Soc. Personal. Psychol. Compass* **2**, 895–908 (2008).
33. Babcock, L., Wang, X. & Loewenstein, G. Choosing the wrong pond: social comparisons in negotiations that reflect a self-serving bias. *Q. J. Econ.* **111**, 1–19 (1996).
34. Mezulis, A. H., Abramson, L. Y., Hyde, J. S. & Hankin, B. L. Is there a universal positivity bias in attributions? A meta-analytic review of individual, developmental, and cultural differences in the self-serving attributional bias. *Psychol. Bull.* **130**, 711–747 (2004).
35. Pronin, E., Lin, D. Y. & Ross, L. The bias blind spot: perceptions of bias in self versus others. *Pers. Soc. Psychol. Bull.* **28**, 369–381 (2002).
36. Regner, T. & Matthey, A. Do reciprocators exploit or resist moral wiggle room? An experimental analysis. *Jena Econ. Res. Pap.* https://econpapers.repec.org/paper/jrpjrpwrp/2015-027.htm (2015).
37. Larson, T. & Capra, C. M. Exploiting moral wiggle room: illusory preference for fairness? A comment. *Judgm. Decis. Mak.* **4**, 467–474 (2009).
38. Dana, J., Weber, R. A. & Kuang, J. X. Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Econ. Theory* **33**, 67–80 (2007).
39. Batson, C. D. Moral masquerades: experimental exploration of the nature of moral motivation. *Phenomenol. Cogn. Sci.* **7**, 51–66 (2008).
40. Babcock, L., Loewenstein, G., Issacharoff, S. & Camerer, C. Biased judgments of fairness in bargaining. *Am. Econ. Rev.* **85**, 1337–1343 (1995).
41. Babcock, L. & Loewenstein, G. Explaining bargaining impasse: the role of self-serving biases. *J. Econ. Perspect.* **11**, 109–126 (1997).
42. Hippel, S. & Hoeppner, S. Biased judgements of fairness in bargaining: a replication in the laboratory. *Int. Rev. Law Econ.* **58**, 63–74 (2019).
43. Farmer, A. & Pecorino, P. Pretrial bargaining with self-serving bias and asymmetric information. *J. Econ. Behav. Organ.* **48**, 163–176 (2002).
44. Wu, J. & Axelrod, R. How to cope with noise in the iterated prisoner's dilemma. *J. Confl. Resolut.* **39**, 183–189 (1995).
45. Boyd, R. *A Different Kind of Animal: How Culture Transformed our Species* (Princeton Univ. Press, 2018).
46. Wiessner, P. Norm enforcement among the Ju/'hoansi bushmen: a case of strong reciprocity? *Hum. Nat.* **16**, 115–145 (2005).
47. Wiessner, P. W. Embers of society: firelight talk among the Ju/'hoansi bushmen. *Proc. Natl Acad. Sci. USA* **111**, 14027–14035 (2014).
48. Mathew, S. & Boyd, R. The cost of cowardice: punitive sentiments towards free riders in Turkana raids. *Evol. Hum. Behav.* **35**, 58–64 (2014).
49. Arno, A. Fijian gossip as adjudication: a communication model of informal social control. *J. Anthropol. Res.* **36**, 343–360 (1980).
50. Merry, S. E. in *Comparative Studies* Vol. 2 (ed. Abel, R. L.) 17–45 (Elsevier, 1982).
51. Vuchinich, S., Emery, R. E. & Cassidy, J. Family members as third parties in dyadic family conflict: strategies, alliances, and outcomes. *Child Dev.* **59**, 1293–1302 (1988).
52. Pearson, J. An evaluation of alternatives to court adjudication. *Justice Syst. J.* **7**, 420–444 (1982).
53. Albert, R. & Howard, D. A. Informal dispute resolution through mediation. *Mediat. Q.* **10**, 99–108 (1985).
54. Heritage, J. & Clayman, S. in *Talk in Action: Interactions, Identities, and Institutions* (eds Heritage, J. & Clayman, S.) 200–212 (Wiley-Blackwell, 2010).
55. Stewart, A. J. & Plotkin, J. B. Collapse of cooperation in evolving games. *Proc. Natl Acad. Sci. USA* **111**, 17558–17563 (2014).
56. Stewart, A. J. & Plotkin, J. B. From extortion to generosity, evolution in the iterated prisoner's dilemma. *Proc. Natl Acad. Sci. USA* **110**, 15348–15353 (2013).
57. Hilbe, C., Nowak, M. A. & Sigmund, K. Evolution of extortion in iterated prisoner's dilemma games. *Proc. Natl Acad. Sci. USA* **110**, 6913–6918 (2013).
58. Osborne, M. J. & Rubinstein, A. *A Course in Game Theory* (MIT Press, 1994).
59. Gurven, M. Reciprocal altruism and food sharing decisions among Hiwi and Ache hunter-gatherers. *Behav. Ecol. Sociobiol.* **56**, 366–380 (2004).
60. Gurven, M., Hill, K., Kaplan, H., Hurtado, A. & Lyles, R. Food transfers among Hiwi foragers of Venezuela: tests of reciprocity. *Hum. Ecol.* **28**, 171–218 (2000).
61. Allen-Arave, W., Gurven, M. & Hill, K. Reciprocal altruism, rather than kin selection, maintains nepotistic food transfers on an Ache reservation. *Evol. Hum. Behav.* **29**, 305–318 (2008).
62. Xue, M. & Silk, J. The role of tracking and tolerance in relationship among friends. *Evol. Hum. Behav.* **33**, 17–25 (2012).
63. Hruschka, D. J. *Friendship: Development, Ecology, and Evolution of a Relationship* (Univ. of California Press, 2010).
64. Stewart-Williams, S. Altruism among kin vs. nonkin: effects of cost of help and reciprocal exchange. *Evol. Hum. Behav.* **28**, 193–198 (2007).
65. Crittenden, A. N. & Zes, D. A. Food sharing among Hadza hunter-gatherer children. *PLoS ONE* **10**, e0131996 (2015).
66. Hammerstein, P. in *Genetic and Cultural Evolution of Cooperation* (ed. Hammerstein, P.) 83–93 (MIT Press, 2003).
67. Clutton-Brock, T. Cooperation between non-kin in animal societies. *Nature* **462**, 51–57 (2009).
68. André, J.-B. Mechanistic constraints and the unlikely evolution of reciprocal cooperation. *J. Evol. Biol.* **27**, 784–795 (2014).
69. Leimar, O. & Hammerstein, P. Cooperation for direct fitness benefits. *Phil. Trans. R. Soc. Lond. B* **365**, 2619–2626 (2010).
70. Raihani, N. J. & Bshary, R. Resolving the iterated prisoner's dilemma: theory and reality. *J. Evol. Biol.* **24**, 1628–1639 (2011).
71. Gilby, I. C. Meat sharing among the Gombe chimpanzees: harassment and reciprocal exchange. *Anim. Behav.* **71**, 953–963 (2006).
72. Watts, D. Reciprocity and interchange in the social relationships of wild male chimpanzees. *Behaviour* **139**, 343–370 (2002).
73. Russell, A. F. & Wright, J. Avian mobbing: byproduct mutualism not reciprocal altruism. *Trends Ecol. Evol.* **24**, 3–5 (2009).
74. von Rohr, C. R. et al. Impartial third-party interventions in captive chimpanzees: a reflection of community concern. *PLoS ONE* **7**, e32494 (2012).
75. Tajima, T. & Kurotori, H. Nonaggressive interventions by third parties in conflicts among captive Bornean orangutans (*Pongo pygmaeus*). *Primates* **51**, 179–182 (2010).
76. Beisner, B. A. & McCowan, B. Policing in nonhuman primates: partial interventions serve a prosocial conflict management function in rhesus macaques. *PLoS ONE* **8**, e77369 (2013).
77. Mathew, S., Boyd, R. & van Veelen, M. in *Cultural Evolution, Strüngmann Forum Report 12* (eds Richerson, P. J. & Christiansen, M.) 45–60 (MIT Press, 2013).

Acknowledgements

We thank M. Hoffman for help with proving that ATFT is subgame perfect. We also thank J. Silk, P. Richerson and J. Henrich for useful comments. This research was funded by the John Templeton Foundation (grant no. 48952). The views expressed in this publication are those of the authors and do not necessarily reflect the views of the funding agency.

Author contributions

S.M. and R.B. conceived the study, developed the model and wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41562-020-01008-1.

Correspondence and requests for materials should be addressed to S.M.

Peer review information Primary handling editors: Charlotte Payne; Mary Elizabeth Sutherland.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): Sarah Mathew

Last updated by author(s): Sep 9, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☒	☐	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection.
Data analysis	Mathematica, Matlab

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Provide your data availability statement here.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	The study uses evolutionary game theory modeling to examine whether third party arbitration can facilitate reciprocal cooperation.
Research sample	No data was collected from subjects. The model was analyzed to examine the parameter conditions under which a strategy that uses third party arbitration can persist.
Sampling strategy	*Describe the sampling procedure (e.g. random, snowball, stratified, convenience). Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient. For qualitative data, please indicate whether data saturation was considered, and what criteria were used to decide that no further sampling was needed.*
Data collection	No data was collected. A mathematical model was developed and analyzed.
Timing	*Indicate the start and stop dates of data collection. If there is a gap between collection periods, state the dates for each sample cohort.*
Data exclusions	*If no data were excluded from the analyses, state so OR if data were excluded, provide the exact number of exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Non-participation	*State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation.*
Randomization	*If participants were not allocated into experimental groups, state so OR describe how participants were allocated to groups, and if allocation was not random, describe how covariates were controlled.*

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging