Synergy between intention recognition and commitments in cooperation dilemmas

The Anh Han[α,⋆], Francisco C. Santos[β,γ], Tom Lenaerts[δ,λ] and Luís Moniz Pereira[ϵ]

January 30, 2015

[α] School of Computing, Teesside University, Borough Road, Middlesbrough, UK TS1 3BA

[β] INESC-ID and Instituto Superior Técnico, Universidade de Lisboa, IST-Taguspark, 2744-016 Porto Salvo, Portugal

[γ] ATP-group, CMAF, Instituto para a Investigação Interdisciplinar, P-1649-003 Lisboa Codex, Portugal

[δ] AI lab, Computer Science Department, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium

[λ] MLG, Département d'Informatique, Université Libre de Bruxelles, Boulevard du Triomphe CP212, 1050 Brussels, Belgium

[ϵ] NOVA Laboratory for Computer Science and Informatics (NOVA LINCS), Departamento de Informática, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal

[⋆] corresponding author: T.Han@tees.ac.uk

Abstract

Commitments have been shown to promote cooperation if, on the one hand, they can be sufficiently enforced, and on the other hand, the cost of arranging them is justified with respect to the benefits of cooperation. When either of these constraints is not met it leads to the prevalence of commitment free-riders, such as those who commit only when someone else pays to arrange the commitments. Here, we show how intention recognition may circumvent such weakness of costly commitments. We describe an evolutionary model, in the context of the one-shot Prisoner's Dilemma, showing that if players first predict the intentions of their co-player and propose a commitment only when they are not confident enough about their prediction, the chances of reaching mutual cooperation are largely enhanced. We find that an advantageous synergy between intention recognition and costly commitments depends strongly on the confidence and accuracy of intention recognition. In general, we observe an intermediate level of confidence threshold leading to the highest evolutionary advantage, showing that neither unconditional use of commitment nor intention recognition can perform optimally. Rather, our results show that arranging commitments is not always desirable, but that they may be also unavoidable depending on the strength of the dilemma.

Introduction

Since Darwin, the problem of explaining the evolution of cooperative behavior has been actively investigated in many fields, from Evolutionary Biology, Ecology, to Economics and Social Science. Several mechanisms responsible for the evolution of cooperation have been proposed, from kin and group selection to direct and indirect reciprocity, to structured population, and to punishment[1–5]. Recently, a large body of economic experiments and theoretical studies have shown that high levels of cooperation can be achieved if reliable agreements can be arranged[6–14]. Arranging prior commitments, such as through enforceable contracts or pledges[8], deposit-refund scheme[11,12] or even emotional or reputation-based commitment devices[7,9], provides incentives for others to cooperate, clarifying the preferences or intentions of others[8,15,16]. However, in human societies, not all cooperative ventures require explicit prior commitments to be made. On the one hand, arranging reliable commitments may be very costly (and take time)[15], which can lead to the prevalence of commitment free-riders, and, on the other hand, others' intentions might be clarified without using a commitment device. Contracts are a popular kind of commitment, which play a key role in enforcing cooperation in modern societies. But even then people occasionally prefer not to rely on using a contract, as are the cases for interactions between relatives or close friends, or between (or with) trustworthy brands. In such cases, partners' cooperative behavior can be envisaged with high confidence. People also do not ask for promise or making threats when partners' motivations can be predicted with high confidence, as doing so may lead to negative reactions or an implication of distrust from them[13,17].

Additionally, human beings are experts in mind reading, particularly at discerning what others are perceiving and intending[18]. An ability to assess intention in others, which is clearly possessed by humans[19,20], has been demonstrated to play a promoting role for the emergence of cooperation. It enables individuals to assess cooperative intention in others in noisy and uncertain environments, and to identify those with an exploitative intent[8,16,21–23]. In addition, behavioral experiments show that people do care about and distinguish between real intentions and outcomes, and that difference plays a crucial role in their decision, for instance, whether to cooperate or to defect, and to reward or to punish[21,24–26]. Although recognizing an intention cannot always be done with high enough confidence to make any decision based on it, an ability to assess intention in others, based on previous experience and available observations at hand, allows choosing cooperative partners even without resorting to commitment devices.

Thus motivated, here we investigate whether a conditional use of commitment through intention recognition can promote the emergence of cooperation in the one-shot Prisoner's Dilemma. In its simple form, a cooperative act (C) is to pay a cost (c) for its co-player to receive a benefit ($b > c$), while a defective act is to spend nothing and thus provides its co-player with no benefit. In a one-shot pairwise interaction, for each player it is better to play D, leading to a zero payoff for both, while both can obtain a higher payoff ($b - c$) if they simultaneously choose C. Here, we consider a strategy, which, at each interaction, attempts first to assess the co-player's intention (whether to cooperate or to defect). Only when it is not confident about what the co-player intends to do in the current interaction, does it propose to the co-player a commitment deal. A commitment proposer pays a cost of arrangement (ϵ) to make the commitment credible, but those who commit but then default have to provide the co-player with a compensation (δ)[27]. It has been shown[11,12,14,27], that substantial levels of cooperation are achieved if both the cost of arranging commitment is small enough compared to the cost of cooperation, and a sufficiently high compensation can be enforced. However, if either of these two conditions is not satisfied, commitment free-riders can take over and become dominant[27]. On the one hand, if the cost of arranging commitment is too large, those who commit and cooperate only if someone else pays to arrange the commitment for them are dominant. On the other hand, when the cost of compensation is too low, for instance due to the difficulty of enforcing the deal afterwards, those who agree on the commitment but then default on it dominate the commitment proposers.

We show that a conditional use of commitments, by means of first assessing intentions of the co-player, can facilitate the commitment free-riding issue, ameliorating the performance of commitment and leading to improved cooperation. The key parameter in our model is a *confidence threshold (θ)*, which is utilized to decide when intention recognition can be relied on (to choose a move), or a commitment deal needs to be arranged to clarify the co-player's intention. The questions we would like to ask here are whether such a conditional use of commitment can resolve the commitment free-riding issues, particularly when a strong commitment cannot be arranged. Furthermore, what is the appropriate confidence threshold, inasmuch the benefit and the cost of commitments and the accuracy of the intention recognition vary?

Results

We consider here, next to the traditional pure cooperator (C) and defector (D) strategies, a new strategy which combines intention recognition and commitment arrangement, denoted by IRCOM. In an interaction, IRCOM recognizes the intention (to cooperate or to defect) of its co-player. A confidence level, $x \in [0, 1]$, is assigned to the recognition result. It defines the degree of confidence, in terms of a probability, that IRCOM predicts the co-player's intention correctly. Then, if it is confident enough about the prediction, that is if x is greater than a given, so-called, *confidence threshold*, $\theta \in [0, 1]$, then in the current interaction it cooperates if the recognized intention of the co-player is to cooperate, and defects otherwise.

When IRCOM is not sufficiently confident about its co-player's intention, i.e. $x < \theta$, it proposes a commitment to others and subsequently cooperates if the opponent accepts the deal. If the deal is not accepted, then this IRCOM refuses to play the game. We consider two additional commitment free-riding strategies[14,27]: (i) The fake committers (FAKE), who accept a commitment proposal yet defect when playing the game, presuming that they can exploit the commitment proposers without suffering a severe consequence; and, (ii) the commitment free-riders (FREE), who defect unless being proposed a commitment, which they then accept and next cooperate in the PD game. In other words, these players are willing to cooperate when a commitment is arranged but are not prepared to pay the cost of setting it up.

However, the prediction being made can be wrong. We assume that prediction accuracy and confidence are positively correlated[28–30]. Namely, the probability of a correct prediction is, $y = r \times x$, where $r > 0$ is dubbed the *accuracy-to-confidence* ratio. Assuming that the confidence, x, are uniformly distributed in $[0, 1]$, the payoff matrix for IRCOM reads

$$M = (1 - \theta)M_1 + \theta M_2, \tag{1}$$

where M_1 and M_2 are the payoff matrices when IRCOM plays without proposing a commitment (i.e. when $x > \theta$) and when it does so (i.e. when $x \leq \theta$), respectively. For details of the computation of the two matrices see Methods and Supporting Information (SI). Table 1 summarizes the parameters and variables in our model.

Note that if $x \leq \theta$, i.e. IRCOM is not confident enough about its intention prediction, it behaves the same as a pure commitment proposer (COMP)[27] when interacting with the non-proposing commitment strategies (i.e. C, D, FAKE and FREE). The greater θ is, the more cautious IRCOM is about its intention recognition result, thereby tending to use commitments more frequently. In an interaction between IRCOM and COMP, we consider that COMP always proposes first and pays the arrangement cost ϵ due to the time delay and effort IRCOM spends on intention recognition deliberation.

Emergence of conditional commitment and cooperation

We first study the stationary distribution in a population of the six above described strategies, namely IRCOM, COMP, C, D, FAKE and FREE (see Methods). The results show that, for a large range of the confidence threshold θ, IRCOM is dominant, whereas the population spends most of the time in the homogenous state of IRCOM, regardless of the initial composition of the population (Figure 1a). However, when θ is low, free-riding strategies become dominant. That is, when IRCOM does not have sufficient confidence about whether its co-player intends to cooperate or to defect in the current interaction, it would be better off counting on arranging a (costly) commitment deal.

Figure 1b shows that the prevalence of IRCOM endures for a wide range of ϵ and δ, as long as an appropriate θ is adopted. Interestingly, in contrast to COMP[27], it is not always the case that the frequency of IRCOM is demolished when ϵ increases (see also Figure S2 in SI). IRCOM actually becomes more frequent when ϵ is sufficiently high, but not too high. This is mainly because IRCOM suppresses the commitment free-riders for a wider range of ϵ, as can be seen from Figure 1d where we show the transition probabilities and the transition directions amongst the six strategies. Namely, for a sufficiently high ϵ (namely, $\epsilon = 2.0$), COMP is taken over by the FREE players, against which IRCOM still is a viable strategy. However, when ϵ is too large, IRCOM is again taken over by FREE players (see Figure S4 in the SI for a larger ϵ). The viability of IRCOM in dealing with commitment free-riders is robust for varying the accuracy-to-confidence ratio, r, as shown in Figure 1c. Namely, we observe that IRCOM is the dominant strategy whenever this ratio is sufficiently high, although the commitment free-riding strategy FREE takes over when r is too small. That is, whenever intention recognition can be performed with a sufficiently high accuracy, as are the case for instance in repeated games[16,23] or when the intention recognition process is facilitated[21,26], IRCOM is amply sufficient at dealing with commitment free-riders.

We now analyze whether and when the conditional use of commitment can actually facilitate the evolution of cooperation. To that end, we make a direct comparison in terms of the level of cooperation obtained through commitment strategies in our model, i.e. from IRCOM and COMP, and such a level in the unconditional commitment model where IRCOM is not included, see Figure 2. The results show that certain improvement is possible for a wide range of commitment deals, i.e. for varying ϵ and δ, see Figure 2a. Interestingly, the improvement is most significant when the commitment deal is weak, that is, when it is rather costly to arrange (high ϵ) and/or no sufficiently high compensation can be enforced (low δ). It is exactly when COMP does not perform well, as it is dominated by the commitment free-riders FREE and FAKE in either condition (i.e. high ϵ or low δ), respectively[27]. This notable observation is robust for varying r, as can be seen in Figure 2b: the improvement in terms of cooperation is positive in general, and increases with r. Furthermore, the improvement is substantial for large ϵ (see for instance cases with $\epsilon = 2$ and 4). In SI, we show that the improvement is also more significant when the benefit-to-cost ratio is larger (see Figure S1).

We now ask, when should one take more risk, avoiding to arrange costly commitment? In Figure 3 we address the effect of varying ϵ and δ, as well as varying the accuracy over confidence ratio r. In general, the higher ϵ and the higher r, the lower confidence level needs to be attained to rely on intention recognition predictions. That is, as the PD becomes more beneficial and the intention recognition prediction can be carried out more accurately, a smaller confidence is exacted to rely on intention recognition, thereby avoiding the cost of arranging commitment. We also observe that this confidence level does not significantly depend on δ, see Figure 3b.

Discussion

We have shown, within the context of the one-shot Prisoner Dilemma (PD), that a conditional use of commitment based on a subjective confidence in assessing a co-player's intention can lead to improved levels of commitment and cooperation. In general, by avoiding the payment of the cost of arranging commitments whenever gaining a sufficient confidence about the co-player's intention, an evolutionary advantage can be achieved. Waiting for a too large confidence may lead to unnecessarily paying the cost, though it can be avoided. However, doing so when confidence is low allows defectors and commitment free-riders to exploit, leading to the destruction of cooperation. Our results show that the gained improvement via the intention recognition capability is more significant when the PD is less harsh, and as more accurate predictions can be achieved. Interestingly, such an improvement is most significant when the cost of arranging commitments is high, thereby overcoming the weaker cases of using the pure commitment strategy[27]. Moreover, our analysis suggests that, as the PD becomes more beneficial and the prediction is more accurate, a smaller confidence is required to enable to take the risk involved in avoiding to arrange costly commitments. These results suggest that, although many societies may have evolved mechanisms to facilitate the making and the enforcement of prior commitments (e.g. legal contracts)[9,15], the cost-efficiency problem faced when implementing such mechanisms (e.g. law systems) may be coped with by using more complex cognitive skills such as of intention recognition (which has been demonstrated to be prevalent in humans and primates[18–20]), in order to facilitate further the sustainability of the commitment mechanisms, hence cooperation.

Our results are in line with the work in[31], where a resource claiming model is described. In that model, players can choose whether to engage in a fight for a resource based on their estimation of the opponents' capability and the players' confidence about their own capacity. It has been shown that overconfidence (which is equivalent to the avoidance of arranging costly commitment at a low confidence threshold in our model) can become evolutionarily stable when the resource is sufficiently large compared to the cost of fighting, as the players might lose their chance of winning the resource if not being confident enough even when they have a stronger capacity than their opponents. Our work differs from this model in that whenever the players have a low confidence level (about their opponents' intention), instead of refusing to play they can make use of the alternative, but provenly efficient strategy, of arranging prior commitments. As we have shown, this combination of the two strategic behaviors performs substantially better than the sole intention recognition one.

The key role of intention recognition in the current model is to allow choosing cooperative partners and avoid reliance on arranging a costly explicit commitment. In environments where partner selection is possible—that is, when people can choose with whom they associate for mutualistic endeavors—then implicit commitments are evolved, by which people behave as if they had bargained with others in order to reach an agreement, in accordance with contractualist moral psychology[32,33]. Hence, our results suggest that intention recognition might have been shaped by natural selection to enable effective partner selection, which in turn drives the evolution of implicit commitments, thereby avoiding the cost of arranging explicit commitments.

Several behavioral experiments on intention based strategies exist that are closely related to our model. The experiment in[26] uses a sequential PD (in the presence of noise) where the second-moving player can recognize the first-moving player's intention, and choose whether to punish a defecting act. The experiment showed that individuals tend to use strong punishment against those who are recognized to have a clear intention of defection while no (or weak) punishment is used against those who defected but the act is recognized to be unintentional. Our work differs from this experimental setting in that the intention recognition process is done prior to the interaction (to find out whether it is necessary to arrange prior commitments), while it is posterior in the experiment, i.e. after the move has been made. Another experiment in[21] showed that, in the course if the repeated Prisoner's Dilemma, if co-players' intention can be observed, it significantly fosters cooperation since unintentional defection caused by noise can be forgiven, as also shown theoretically in[22]. Note that both experiments have been designed so that the intention recognition process is facilitated, thereby guaranteeing a high confidence level. In such cases, as shown in the present work, the synergy of intention recognition and commitments, both aiming at clarifying co-players' intention, can promote a high level of cooperation.

Several extensions to the present model can be described. In our model we have considered a general one-shot interaction scenario, but we envisage that as more prior experience is incorporated, for instance by observing direct or indirect past actions of the co-player, intention recognition can be performed better, thereby leading to better performance of IRCOM. Indeed, in[22,34], in the context of the repeated PD with implementation noise, Artificial Intelligence based intention recognition strategies[35,36] can more accurately assess a co-player's intention whenever more past interactions are taken into account. In SI, we consider a more effective IRCOM strategy, having a more accurate intention recognition capability (see Figure S3). Our numerical results show that, whenever the intention recognition model is efficient enough, the intention recognition strategy by itself alone (i.e. IRCOM with $\theta = 0$) performs quite well, complying with the results obtained in[22,34], where concrete intention recognition models are deployed.

Overall, our work indicates that, on the one hand, it is evolutionarily advantageous to be able to avoid arranging costly commitments whenever the co-player's intention can be assessed with sufficient confidence and accuracy. On the other hand, arranging prior commitments may be also unavoidable, depending on the strength of the dilemma, in order to reach a high level of cooperation.

Methods

Our analysis is based on evolutionary game theory methods for finite populations[37,38]. In the context of evolutionary game theory, the individuals' or agents' payoff represents their *fitness* or social *success*. The dynamics of strategy change in a population is governed by social learning, that is, the most successful agents will tend to be imitated by the others. There are many ways to model social learning[5,39,40]. Adopting one of the most frequently used ones, we consider the so-called pairwise comparison rule[41], which assumes that an agent A with fitness f_A adopts the strategy of another agent B with fitness f_B with probability given by

$$\frac{1}{1+e^{-\beta(f_B-f_A)}},$$

where β controls the 'imitation strength', i.e., how strongly the agents are basing the decision to imitate on fitness comparisons. For $\beta = 0$, we obtain the limit of neutral drift – the imitation decision is random. For large β, imitation becomes increasingly deterministic.

In the absence of mutations, the end states of evolution are inevitably monomorphic: once such a state is reached, imitation cannot produce any change. We thus further assume that, with a certain mutation probability $\mu > 0$ (also dubbed the exploration rate[42]), an agent switches randomly to a different strategy without imitating any other agent. The resulting Markov Chain has a stationary distribution, which characterizes the average time the population spends in each of these monomorphic end states. Yet, for arbitrary exploration rates and number of strategies, stationary distributions are often cumbersome to compute[43–45].

Fortunately, in the case of small exploration or mutation rates, analytical computation of this stationary distribution can conveniently be computed[38,43,46,47]. The small exploration rates guarantee that any newly occurred mutant in a homogeneous population will fixate or become extinct long before the occurrence of another mutation. Hence, the population will always consist of at most two strategies in co-presence. This allows one to describe the evolutionary dynamics of our population in terms of a reduced Markov Chain, whose size is equal the number of strategies being considered, and each state represents a possible monomorphic end state of the population associated with a one of the strategies. The transitions between states are defined by the fixation probabilities of a single mutant of one strategy in a homogeneous population of individuals adopting another strategy[46].

More precisely, let N be the size of the population. Suppose there are at most two strategies in the population, say, k agents using strategy A ($0 \leq k \leq N$) and $(N-k)$ agents using strategy B. Thus, the (average) payoff of the agent that uses A or uses B can be written as follows, respectively,

$$\begin{aligned} \Pi_A(k) &= \frac{(k-1)\pi_{A,A} + (N-k)\pi_{A,B}}{N-1}, \\ \Pi_B(k) &= \frac{k\pi_{B,A} + (N-k-1)\pi_{B,B}}{N-1}, \end{aligned} \tag{2}$$

where $\pi_{X,Y}$ stands for the payoff an agent using strategy X obtained in an interaction with another agent using strategy Y, given by the payoff matrix (9).

Now, the probability to change, by ± 1, the number k of agents using strategy A at each time step can be written as

$$T^{\pm}(k) = \frac{N-k}{N}\frac{k}{N}\frac{1}{1+e^{\mp\beta[\Pi_A(k)-\Pi_B(k)]}}. \tag{3}$$

The fixation probability of a single mutant with a strategy A in a population of $(N-1)$ agents using B is given by[38,41,43,46,48]

$$\rho_{B,A} = \frac{1}{1+\sum_{i=1}^{N-1}\prod_{j=1}^{i}\frac{T^-(j)}{T^+(j)}}. \tag{4}$$

In the limit of neutral selection ($\beta = 0$), $T^{-}(j) = T^{+}(j)\ \forall j$. Thus, $\rho_{B,A} = 1/N$. Considering a set $\{1, ..., q\}$ of different strategies, these fixation probabilities determine a transition matrix $M = \{T_{ij}\}_{i,j=1}^{q}$, with $T_{ij,j\neq i} = \rho_{ji}/(q-1)$ and $T_{ii} = 1 - \sum_{j=1,j\neq i}^{q} T_{ij}$, of a Markov Chain. The normalized eigenvector associated with the eigenvalue 1 of the transposed of M provides the stationary distribution described above[38,43,46,48], describing the relative time the population spends adopting each of the strategies.

Deriving Payoff Matrix The one-shot Prisoner's Dilemma can be described with the following payoff matrix:

$$\begin{array}{cc} & \begin{array}{cc} C & D \end{array} \\ \begin{array}{c} C \\ D \end{array} & \begin{pmatrix} R, R & S, T \\ T, S & P, P \end{pmatrix} \end{array}.$$

Once the interaction is established and both players have decided to play C or D (with or without commitment arrangements), both players receive the same reward R (penalty P) for mutual cooperation (mutual defection). Unilateral cooperation provides the sucker's payoff S for the cooperative player and the temptation to defect T for the defecting one. The payoff matrix corresponds to the preferences associated with the Prisoner's Dilemma when the parameters satisfy the ordering, $T > R > P > S$[5,49]. In the main text, we use the Donor game, a special case of the PD, with $T = b$; $R = b - c$; $P = 0$; $S = -1$, where b and c are the benefit and cost of cooperation, respectively.

When proposing commitment, the average payoff of IRCOM, as the row player, reads[27]

$$M_1 = \begin{array}{c} \\ COMP \\ C \\ D \\ FAKE \\ FREE \end{array} \begin{array}{c} \begin{array}{ccccc} COMP & C & D & FAKE & FREE \end{array} \\ \begin{pmatrix} R-\epsilon/2 & R-\epsilon & 0 & S+\delta-\epsilon & R-\epsilon \\ R & R & S & S & S \\ 0 & T & P & P & P \\ T-\delta & T & P & P & P \\ R & T & P & P & P \end{pmatrix} \end{array}. \tag{5}$$

The probability that IRCOM relies on the intention recognition prediction, and the prediction was actually correct, can be written as joint probability distribution[50]

$$p_c = P(x > \theta, y < \min\{rx, 1\}) = \int_\theta^{+\infty}\int_0^{\min\{rx,1\}} dy\, dx = \begin{cases} \frac{r(1-\theta)(1+\theta)}{2} & \text{if } r \leq 1 \text{ or } r \geq \frac{1}{\theta} \\ 1 - \frac{1}{2r} - \frac{r\theta^2}{2} & \text{otherwise .} \end{cases} \tag{6}$$

Similarly, the probability that IRCOM relies on the intention recognition prediction, but the prediction was not correct, is

$$p_{ic} = \begin{cases} (1-\theta)\left[1 - \frac{r(1+\theta)}{2}\right] & \text{if } r \leq 1 \text{ or } r \geq \frac{1}{\theta}, \\ \frac{1}{2r} + \frac{r\theta^2}{2} - \theta & \text{otherwise.} \end{cases} \tag{7}$$

Hence, IRCOM cooperation probability when playing with another IRCOM player is, $\theta + p_c$.

The payoff matrix for IRCOM when relying on intention recognition reads

$$M_2 = (1-\theta)\begin{pmatrix} m_{11} & m_{12} & m_{13} & m_{14} & m_{15} \\ m_{21} & R & S & S & S \\ m_{31} & T & P & P & P \\ m_{41} & T & P & P & P \\ m_{51} & T & P & P & P \end{pmatrix}, \tag{8}$$

where

$m_{11} = Pu^2 + (S+T)uv + Rv^2$;

$m_{12} = Tu + Rv$;

$m_{15} = m_{14} = m_{13} = Su + Pv$;

$m_{21} = Su + Rv$;

$m_{51} = m_{41} = m_{31} = Tu + Pv$;

With $u = p_{ic}/(1-\theta)$ and $v = p_c/(1-\theta)$.

Finally, the payoff matrix for IRCOM (as a row player) reads

$$M = (1-\theta)M_1 + \theta M_2. \tag{9}$$

References

1. Nowak, M. A. Five rules for the evolution of cooperation. *Science* **314**, 1560 (2006). DOI: 10.1126/science.1133755.

2. Nowak, M. A. *Evolutionary Dynamics: Exploring the Equations of Life* (Harvard University Press, Cambridge, MA, 2006).

3. West, S., Griffin, A. & Gardner, A. Evolutionary explanations for cooperation. *CURR BIOL* **17**, R661–R672 (2007).

4. Perc, M., Gómez-Gardeñes, J., Szolnoki, A., Floría, L. M. & Moreno, Y. Evolutionary dynamics of group interactions on structured populations: a review. *J R SOC INTERFACE* **10**, 20120997 (2013).

5. Sigmund, K. *The Calculus of Selfishness* (Princeton University Press, 2010).

6. Ostrom, E. *Governing the commons: The evolution of institutions for collective action* (Cambridge university press, 1990).

7. Frank, R. H. *Passions Within Reason: The Strategic Role of the Emotions* (W. W. Norton and Company, 1988).

8. Chen, X.-P. & Komorita, S. S. The effects of communication and commitment in a public goods social dilemma. *Organ Behav Hum Decis Process* **60**, 367–386 (1994).

9. Nesse, R. M. *Evolution and the capacity for commitment*. Russell Sage Foundation series on trust (Russell Sage, 2001).

10. Goodenough, O. R. Law and the biology of commitment. In Nesse, R. M. (ed.) *Evolution and the capacity for commitment*, 262–291 (New York: Russell Sage, 2001).

11. Gerber, A. & Wichardt, P. C. Providing public goods in the absence of strong institutions. *J PUBLIC ECON* **93**, 429–439 (2009).

12. Cherry, T. L. & McEvoy, D. M. Enforcing compliance with environmental agreements in the absence of strong institutions: An experimental analysis. *Environ Resour Econ* **54**, 63–77 (2013).

13. Miettinen, T. Promises and conventions–an approach to pre-play agreements. *Games Econ Behav* **80**, 68–84 (2013).

14. Han, T. A., Pereira, L. M. & Lenaerts, T. Avoiding or Restricting Defectors in Public Goods Games? *J R SOC INTERFACE* **12**, 20141203 (2014).

15. Sterelny, K. *The evolved apprentice* (MIT Press, 2012).

16. Han, T. A. *Intention Recognition, Commitments and Their Roles in the Evolution of Cooperation: From Artificial Intelligence Techniques to Evolutionary Game Theory Models*, vol. 9 (Springer SAPERE series, 2013).

17. Frank, R. H. Cooperation through Emotional Commitment. In Nesse, R. M. (ed.) *Evolution and the capacity for commitment*, 55–76 (New York: Russell Sage, 2001).

18. Tomasello, M., Carpenter, M., Call, J., Behne, T. & Moll, H. Understanding and sharing intentions: The origins of cultural cognition. *Behav. Brain. Sci.* **28**, 675–691 (2005).

19. Meltzoff, A. N. The framework for recognizing and becoming an intentional agent. *Acta Psychol Amst* **124**, 26–43 (2007).

20. Woodward, A. L., Sommerville, J. A., Gerson, S., Henderson, A. M. & Buresh, J. The emergence of intention attribution in infancy. In Ross, B. H. (ed.) *The Psychology of*

Learning and Motivation, vol. 51 of *Psychology of Learning and Motivation*, 187 – 222 (Academic Press, 2009).

21. Rand, D. G., Fudenberg, D. & Dreber, A. It's the thought that counts: The role of intentions in reciprocal altruism. *Available at SSRN 2259407* (2013).

22. Han, T. A., Pereira, L. M. & Santos, F. C. Intention recognition promotes the emergence of cooperation. *Adapt. Behav.* **19**, 264–279 (2011).

23. Han, T. A., Pereira, L. M. & Santos, F. C. Corpus-based intention recognition in cooperation dilemmas. *Artif. Life* **18**, 365–383 (2012).

24. Charness, G. & Levine, D. I. Intention and stochastic outcomes: An experimental study. *ECON J* **117**, 1051–1072 (2007).

25. Falk, A., Fehr, E. & Fischbacher, U. Testing theories of fairness—intentions matter. *Games Econ Behav* **62**, 287 – 303 (2008).

26. Fischbacher, U. & Utikal, V. On the acceptance of apologies. *Games Econ Behav* **82**, 592–608 (2013).

27. Han, T. A., Pereira, L. M., Santos, F. C. & Lenaerts, T. Good agreements make good friends. *Sci. Rep.* **3** (2013).

28. Jersild, A. The determinants of confidence. *AM J COMMUN PSYCHOL* 640–642 (1929).

29. Busey, T. A., Tunnicliff, J., Loftus, G. R. & Loftus, E. F. Accounts of the confidence-accuracy relation in recognition memory. *Psychon Bull Rev* **7**, 26–48 (2000).

30. Pallier, G. *et al.* The role of individual differences in the accuracy of confidence judgments. *J GEN PSYCHOL* **129**, 257–299 (2002).

31. Johnson, D. D. P. & Fowler, J. H. The evolution of overconfidence. *Nature* **477**, 317–320 (2011).

32. Baumard, N. & Boyer, P. Explaining moral religions. *TRENDS COGN SCI* **17**, 272–280 (2013).

33. Baumard, N., André, J.-B. & Sperber, D. A mutualistic approach to morality: The evolution of fairness by partner choice. *Behav. Brain. Sci.* **36**, 59–78 (2013).

34. Han, T. A., Pereira, L. M. & Santos, F. C. The role of intention recognition in the evolution of cooperative behavior. In Walsh, T. (ed.) *Proceedings of the 22nd international joint conference on Artificial intelligence (IJCAI'2011)*, 1684–1689 (AAAI, 2011).

35. Han, T. A. & Pereira, L. M. State-of-the-art of intention recognition and its use in decision making – a research summary. *AI COMMUN* **26**, 237–246 (2013).

36. Han, T. A. & Pereira, L. M. Context-dependent incremental decision making scrutinizing the intentions of others via bayesian network model construction. *Intelligent Decision Technologies* **7**, 293–317 (2013).

37. Nowak, M. A., Sasaki, A., Taylor, C. & Fudenberg, D. Emergence of cooperation and evolutionary stability in finite populations. *Nature* **428**, 646–650 (2004).

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

39. Hofbauer, J. & Sigmund, K. *Evolutionary Games and Population Dynamics* (Cambridge University Press, 1998).

40. Rendell, L. *et al.* Why copy others? insights from the social learning strategies tournament. *Science* **328**, 208–213 (2010). URL http://www.sciencemag.org/content/328/5975/208.abstract. http://www.sciencemag.org/content/328/5975/208.full.pdf.

41. Traulsen, A., Nowak, M. A. & Pacheco, J. M. Stochastic dynamics of invasion and fixation. *Phys. Rev. E* **74**, 11909 (2006).

42. Traulsen, A., Hauert, C., De Silva, H., Nowak, M. A. & Sigmund, K. Exploration dynamics in evolutionary games. *Proc. Natl. Acad. Sci. USA* **106**, 709–712 (2009).

43. Hauert, C., Traulsen, A., Brandt, H., Nowak, M. A. & Sigmund, K. Via freedom to coercion: The emergence of costly punishment. *Science* **316**, 1905–1907 (2007).

44. Sigmund, K., Silva, H. D., Traulsen, A. & Hauert, C. Social learning promotes institutions for governing the commons. *Nature* **466**, 7308 (2010).

45. Santos, F. C. & Pacheco, J. M. Risk of collective failure provides an escape from the tragedy of the commons. *Proc. Natl. Acad. Sci. USA* **108**, 10421–10425 (2011).

46. Fudenberg, D. & Imhof, L. A. Imitation processes with small mutations. *J ECON THEORY* **131**, 251–262 (2005).

47. Santos, F. C., Pacheco, J. M. & Skyrms, B. Co-evolution of pre-play signaling and cooperation. *J. Theor. Biol.* **274**, 30–35 (2011).

48. Karlin, S. & Taylor, H. E. *A First Course in Stochastic Processes* (Academic Press, New York, 1975).

49. Coombs, C. H. A reparameterization of the prisoner's dilemma game. *Behav. Sci.* **18**, 424–428 (1973).

50. Gut, A. *An Intermediate Course in Probability* (Springer Publishing Company, Incorporated, 2009), 2nd edn.

Acknowledgements

T.A.H acknowledges the support provided by the F.W.O. Belgium. T.L. acknowledges the support provided by the F.R.S. - F.N.R.S Belgium and the F.W.O. Belgium. F.C.S acknowledges the support provided by FCT-Portugal.

Author Contributions

T.A.H., F.C.S., T.L. and L.M.P. designed the research. The models were implemented by T.A.H. Results were analyzed and improved by T.A.H., F.C.S., T.L. and L.M.P. T.A.H., F.C.S., T.L. and L.M.P. wrote the paper together.

Competing Financial Interests

none

Table 1. Variables and parameters used in the model.

Symbols	Description
ϵ	The cost of arranging a commitment deal
δ	The compensation cost
c	The cost of cooperation in the PD game
b	The benefit of cooperation in the PD game
x	The degree of confidence in a correct intention prediction
θ	The confidence threshold to rely on intention recognition (i.e. if $x > \theta$)
r	The accuracy-to-confidence ratio
y	The accuracy of intention prediction, given the confidence ($y = r \times x$)
β	The intensity of selection

394 Figure Legends

Figure 1. (a) Frequency of each strategy as a function of confidence threshold θ. In a population of IRCOM, COMP, C, D, FAKE and FREE individuals, for a sufficiently large θ, IRCOM is most frequent in the population. The performance of IRCOM decreases when θ is too high. It implies that IRCOM should not be too cautious about its intention recognition capacity, i.e. not be too careful to always propose commitment instead of believing in its prediction accuracy; **(b) Frequency of IRCOM at the optimal confidence threshold, as a function of the cost of arranging commitment ϵ and the compensation cost δ.** Interestingly, in contrast to COMP, it is not always the case that the frequency of IRCOM is smaller for larger ϵ. IRCOM is actually more frequent when ϵ is sufficiently large. **(c) Frequency of each strategy as a function of accuracy to confidence ratio, r, at the optimal confidence threshold.** When intention recognition accuracy is sufficiently high, IRCOM is prevalent, but when it is small, FREE is most abundant. **(d) Transitions probabilities and stationary distributions** ($\theta = 0.28$). Note the transitions from COMP to FREE to IRCOM. For clarity, only the transitions that are larger than neutral are shown ($\rho_N = 1/N$ denotes the neutral transition probability). Parameters: In panels (a), (c) and (d): $\delta = 4; \epsilon = 2$; In panels (a), (b) and (d): $r = 1$; In all cases, $b = 4,\ c = 1$; $N = 100$; $\beta = 0.1$.

Figure 2. (a) Improvement in cooperation level obtained from IRCOM and COMP compared to the case where there is no IRCOM, as a function of the cost of arranging commitment ϵ and the compensation cost δ. Improvement is achieved for a wide range of ϵ and δ. It is most significant when ϵ is rather high and δ is not too large, i.e. the commitment deal is weak (see Figure S1 in SI for the improvement obtained in percentage, and also for other parameter values). **(b) Such improvement as a function of the accuracy-to-confidence ration, r, and for different commitment deals**. In general, the larger r, the more significant improvement is obtained. Furthermore, when r is sufficiently high, larger improvement is obtained when it is costly to arrange commitments and/or a high compensation is difficult to enforced. Parameters: $b = 4$, $c = 1$, $N = 100$, and $\beta = 0.1$.

Figure 3. Optimal confidence threshold, (a) as a function of r, for different commitment deals, and (b) as a function of ϵ and δ. In general, the higher r and the larger ϵ, the lower confidence level needs to be attained to rely on intention recognition predictions (i.e. taking higher risk). This confidence level does not significantly depend on δ. We adopt, in both cases, $b = 4,\ c = 1$, $N = 100$, and $\beta = 0.1$. In panel (b), r = 1.

Supporting Information: Synergy between intention recognition and commitments in cooperation dilemmas

The Anh Han[α], Francisco C. Santos[β,γ], Tom Lenaerts[δ,λ] and Luís Moniz Pereira[ϵ]

January 19, 2015

[α] School of Computing, Teesside University, Borough Road, Middlesbrough, UK TS1 3BA

[β] INESC-ID and Instituto Superior Técnico, Universidade de Lisboa, IST-Taguspark, 2744-016 Porto Salvo, Portugal

[γ] ATP-group, CMAF, Instituto para a Investigação Interdisciplinar, P-1649-003 Lisboa Codex, Portugal

[δ] AI lab, Computer Science Department, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium

[λ] MLG, Département d'Informatique, Université Libre de Bruxelles, Boulevard du Triomphe CP212, 1050 Brussels, Belgium

[ϵ] NOVA Laboratory for Computer Science and Informatics (NOVA LINCS), Departamento de Informática, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal

*corresponding author: T.Han@tees.ac.uk

In this supporting information, we provide additional numerical results to show the robustness of our conclusions in the main text.

1 Results for different benefit-to-cost ratios

In Figure S1 we show the cooperation level from commitment strategies, IRCOM and COMP, as a function of the cost of arranging commitment ϵ and the compensation cost δ, the improvement in cooperation level compared to the case where there is no IRCOM, and such an improvement in percentage. We also plot the same quantity for different b/c. In general, we observe that improvement is always possible, and furthermore, the larger b/c (i.e. the less harsh the PD), the larger the improvement is achieved.

Figure S2 shows the frequency of COMP and IRCOM (at the optimal confidence threshold) for different values of ϵ and δ, and for different b/c ratios. In general, for sufficiently large δ and low ϵ, IRCOM dominates the population. Interestingly, in contrast to COMP, it is not always the case that the frequency of IRCOM is smaller for larger ϵ. IRCOM is actually more frequent when ϵ is sufficiently high, which is larger for larger b/c.

2 More efficient intention recognition

In the main text we have used a very inefficient intention recognition model, where the accuracy of intention recognition is a random number derived from $[0, 1]$. It is not surprising that the performance of the intention recognition strategy solely—which corresponds to IRCOM with $\theta = 0$, is very poor. In the sequel, let us study the model using more efficient intention recognition models (Figure S3).

We consider that the prediction accuracy, Y, is randomly distributed in the interval $[\gamma, 1]$, where a larger γ reflects a more efficient intention recognizer at work. In an increasing order of efficiency, Y is uniformly drawn from intervals $[0, 1]$, $[0.3, 1]$, $[0.6, 1]$, and $[0.9, 1]$. Note that in the context of iterated interactions (e.g. in the framework of the iterated Prisoner's Dilemma), these levels of efficiency can be achieved (on average) by considering large enough numbers of

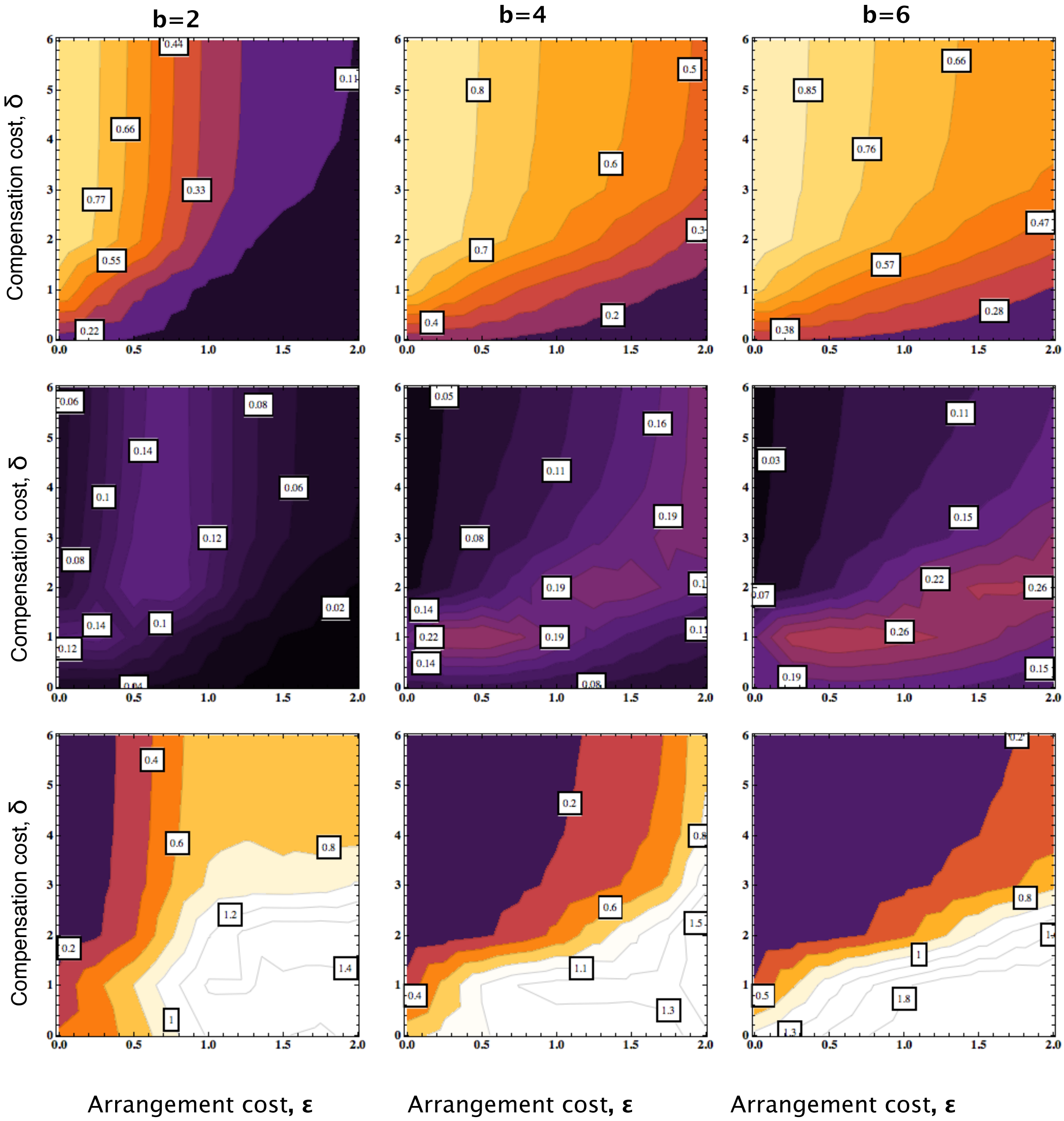

Figure S1: Cooperation level from commitment strategies, IRCOM and COMP, as a function of the cost of arranging commitment ϵ and the compensation cost δ (first row); Improvement in cooperation level compared to the case where there is no IRCOM (second row); and such an improvement in percentage (third row). We plot for different b/c. The larger b/c, the larger the improvement. Parameters: Panels (a) and (b): $\delta = 4$; $\epsilon = 0.7$; In all cases, $b = 2,\ c = 1$; $r = 1$; $N = 100$; $\beta = 0.1$.

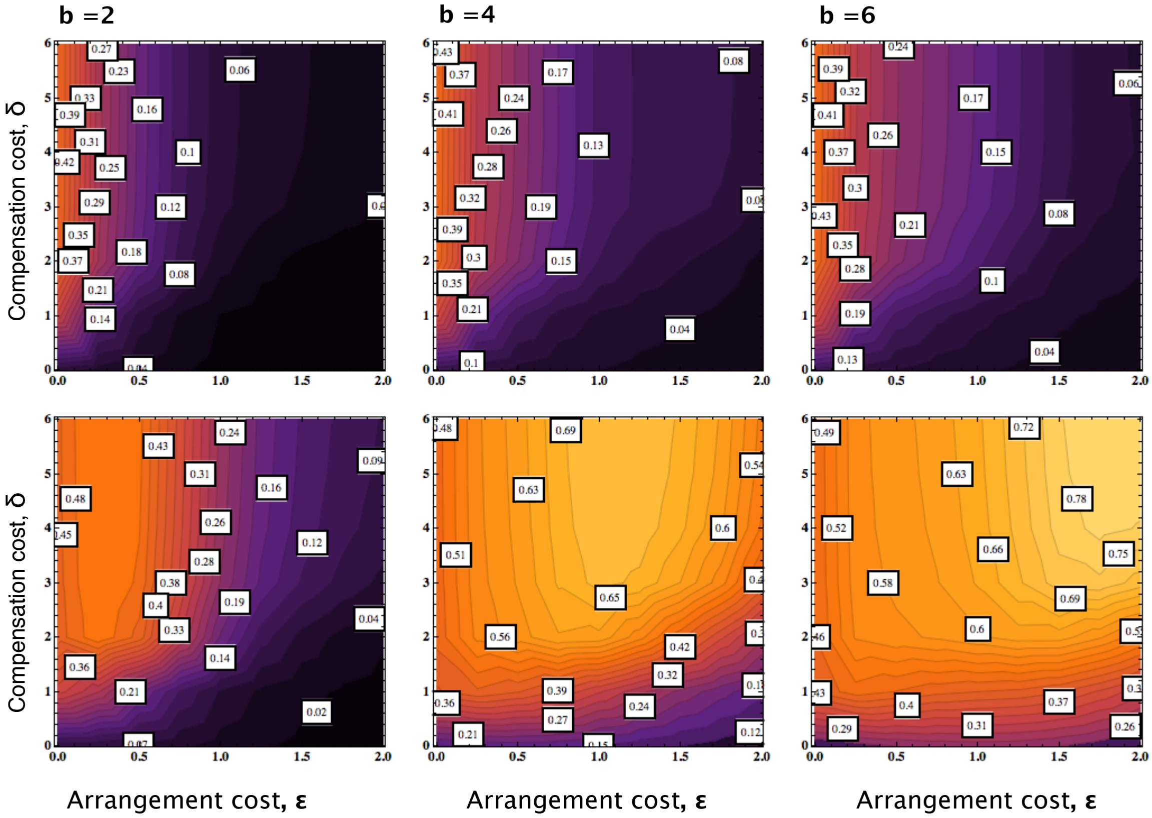

Figure S2: Frequency COMP in a population of five strategies COMP, C, D, FREE, and FAKE (top row) and of IRCOM (at optimal confidence threshold) in a population of six strategies IRCOM, COMP, C, D, FREE, and FAKE (bottom row) for varying ϵ and δ, and for different b/c. In general, for sufficiently large δ, IRCOM dominates the population for small ϵ. Interestingly, in contrast to COMP, it is not always the case that the frequency of IRCOM is smaller for larger ϵ. IRCOM is actually more frequent when ϵ is sufficiently large, which is larger for larger b/c. Parameters: In all cases: $r = 1$; $N = 100$; $\beta = 0.1$.

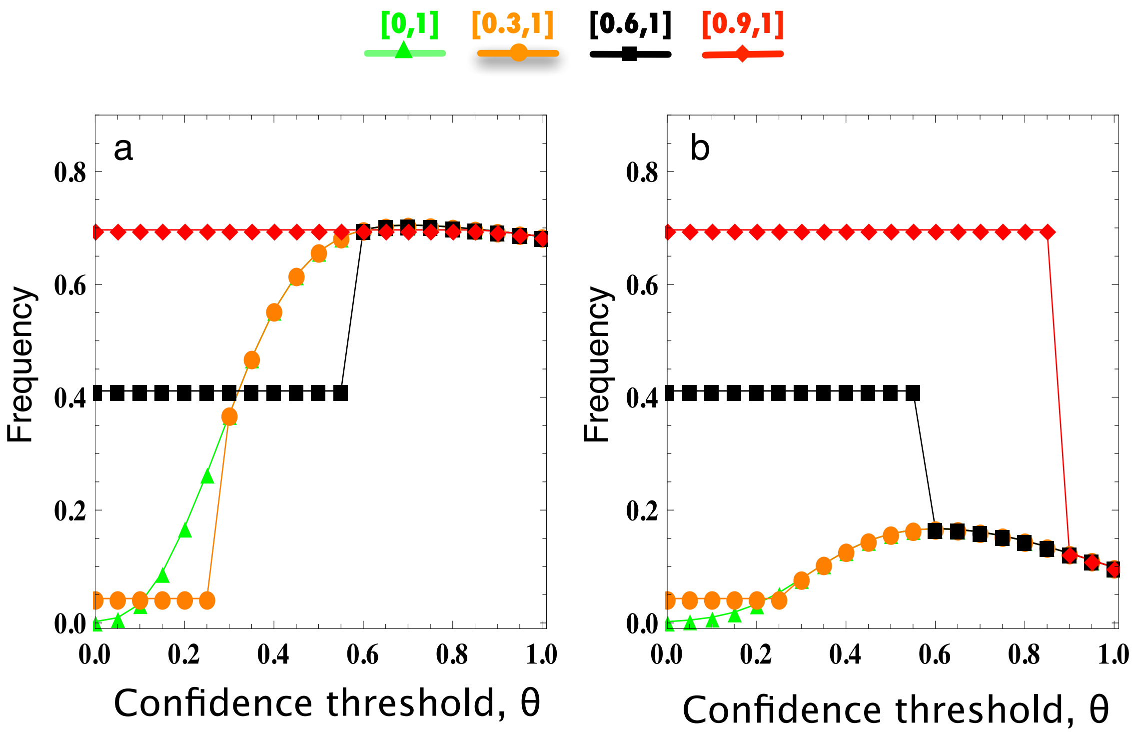

Figure S3: Frequency of IRCOM as a function of confidence threshold, θ, in a population of IRCOM, C, D, FAKE and FREE individuals. We consider different probability distributions of the intention prediction accuracy, reflecting the efficiency or precision of the intention recognition model at work. Namely, Y is uniformly drawn from $[\gamma, 1]$, with $\gamma = 0, 0.3, 0.6, 0.9$. The results show that when intention recognition is highly accurate, it is worth relying more on the intention predictions, even exclusively (see $\gamma = 0.9$ in panel a, and $\gamma = 0.6$ and 0.9 in panel b). Parameters: $\epsilon = 0.25,\ \delta = 4$ (panel a) and $\epsilon = 1,\ \delta = 2$ (panel b); payoff entries, $T = 2,\ R = 1,\ P = 0,\ S = -1$; accuracy over confidence ratio, $r = 1$; population size, $N = 100$; imitation strength, $\beta = 0.1$.

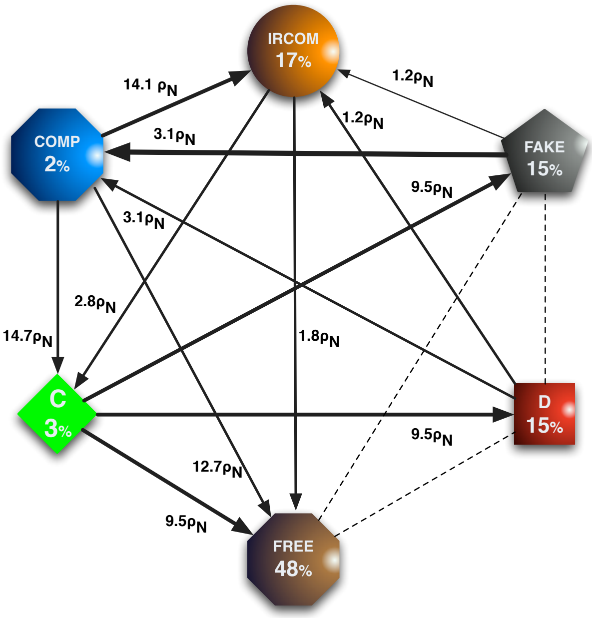

Figure S4: Transitions probabilities and stationary distributions for a large ϵ ($\epsilon = 3$). Other parameters similar to main text: $\delta = 4, r = 1; b = 4,\ c = 1; N = 100; \beta = 0.1$.

interactions between two players (or high enough probabilities of a next interaction[4,5]), given that the noise is small enough. Normally, the more an intention recognizer interacts with a fixed co-player, the better it predicts its co-player's intention. For example, this holds for the two intention recognition models described in[2,3]. Furthermore, in[1], the authors present experimental evidence showing that, in a one-shot PD, subjects of only brief acquaintance were able to recognize players with an intention to defect with more than twice chance accuracy.

The results show that, whenever the intention recognition model is efficient enough, the intention recognition strategy solely (i.e. IRCOM with $\theta = 0$) performs quite well, complying with the results obtained in[2,3], where concrete intention recognition models are deployed. However, when a quite strong commitment deal can be envisaged (Figure S3a), arranging it can still glean some evolutionary advantage. But in case that only weak commitment deals can be arranged (Figure S3b), it is then more beneficial to rely, even exclusively, on the intention recognition strategy should it be efficient enough.

References

[1] Frank, R. H., Gilovich, T., and Regan, D. T. 1993. The evolution of one-shot cooperation: An experiment. *Ethology and Sociobiology*, 14(4):247 – 256.

[2] Han, T. A., Pereira, L. M., and Santos, F. C. 2011. Intention recognition promotes the emergence of cooperation. *Adaptive Behavior*, 19(3):264–279.

[3] Han, T. A., Pereira, L. M., and Santos, F. C. 2012. Corpus-based intention recognition in cooperation dilemmas. *Artificial Life journal*, 18(4):365–383.

[4] Nowak, M. A. 2006. Five rules for the evolution of cooperation. *Science*, 314(5805):1560. DOI: 10.1126/science.1133755.

[5] Sigmund, K. 2010. *The Calculus of Selfishness*. Princeton University Press.