Complex Systems of Mindful Entities: On Intention Recognition and Commitment

Luís Moniz Pereira, The Anh Han and Francisco C. Santos

Abstract The mechanisms of emergence and evolution of cooperation in populations of abstract individuals with diverse behavioural strategies in co-presence have been undergoing mathematical study via Evolutionary Game Theory, inspired in part on Evolutionary Psychology. Their systematic study resorts as well to implementation and simulation techniques, thus enabling the study of aforesaid mechanisms under a variety of conditions, parameters, and alternative virtual games. The theoretical and experimental results have continually been surprising, rewarding, and promising. Recently, in our own work we have initiated the introduction, in such groups of individuals, of cognitive abilities inspired on techniques and theories of Artificial Intelligence, namely those pertaining to both Intention Recognition and to Commitment (separately and jointly), encompassing errors in decision-making and communication noise. As a result, both the emergence and stability of cooperation become reinforced comparatively to the absence of such cognitive abilities. This holds separately for Intention Recognition and for Commitment, and even more when they are engaged jointly. The present paper aims to sensitize the reader to these Evolutionary Game Theory based studies and issues, which are accruing in importance for the modelling of

L. M. Pereira (✉) · T. A. Han
Centro de Inteligência Artificial (CENTRIA), Departamento de Informática,
Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa,
2829-516 Caparica, Portugal
e-mail: lmp@fct.unl.pt

T. A. Han
e-mail: h.anh@ai.vub.ac.be

T. A. Han
AI-lab, Vrije Universiteit Brussel, Pleinlaan 2 1050 Brussels, Belgium

F. C. Santos
INESC-ID, Instituto Superior Técnico and ATP-group, Instituto para
a Investigação Interdisciplinar, Universidade Técnica de Lisboa, IST-Taguspark
2744-016 Porto Salvo, Portugal
e-mail: franciscocsantos@ist.utl.pt

L. Magnani (ed.), *Model-Based Reasoning in Science and Technology*,
Studies in Applied Philosophy, Epistemology and Rational Ethics 8,
DOI: 10.1007/978-3-642-37428-9_28, © Springer-Verlag Berlin Heidelberg 2014

minds with machines, with impact on our understanding of the evolution of mutual tolerance and cooperation. In doing so, it also provides a coherent bird's-eye view of our own varied recent work, whose more technical details and results are spread throughout a number of well recognized publishing venues, and to which we refer the reader for a fuller support of our claims where felt necessary.

1 Introduction

Biological evolution is characterized by a set of highly braided processes, which produce a kind of extraordinarily complex combinatorial innovation. A generic term frequently used to describe this vast category of spontaneous, and weakly predictable, order generating processes, is "emergence". This term became a kind of signal to refer the paradigms of research sensitive to systemic factors. Complex dynamic systems can spontaneously assume patterns of ordered behaviours which are not previously imaginable from the properties of their composing elements nor from their interaction patterns. There is unpredictability in self-organizing phenomena—preferably called *evolutionary*—, with considerably diverse and variable levels of complexity. What does emerge? The answer is not something pre-defined but instead something like a shape, pattern, or function. The concept of emergence is applicable to phenomena in which the relational properties predominate over the properties of composing elements in the determination of the ensemble's characteristics. Emergence processes appear due to configurations and topologies, not to intrinsic properties of elements [16].

The problem of evolution of cooperation and of the emergence of collective action—cutting across areas as diverse as Biology, Economy, Artificial Intelligence, Political Science, or Psychology—is one of the greatest interdisciplinary challenges science faces today [3, 37, 64, 94]. To understand the evolutionary mechanisms that promote and keep cooperative behaviour is all the more complex as increasingly intricate is the intrinsic complexity of the partaking individuals. *Complexity* refers to the study of the emergence of collective properties in systems with many interdependent components. These components can be atoms or macromolecules in a physical or biological context, and people, machines or organizations in a socioeconomic context.

Egotism concerns the logic behind the unending give-and-take that pervades our societal lives. It does not mean blind greed, but instead an informed individual interest. Thus, the evolution of cooperation has been considered one of the most challenging problems of the century. Throughout the ages, the issue of self-consideration versus "the other"-consideration has fascinated thinkers, but the use of formal models and experimental games is relatively recent. Since Robert Trivers [104, 105] introduced the evolutionary approach to reciprocity, games have served as models to explore the issue. The modelling of artificial societies based on the individual has significantly expanded the scope of game theory. Societies are composed by fictitious subjects, each equipped with a strategy specified by a program. Individuals repeatedly meet other individuals, each time doing so in randomized pairs, in a joint iterated game taking place within the scope of the whole population. The comparison of accumulated rewards is used to update the population: the most successful individuals produce more offspring, which inherit their strategy. Alternatively, instead of inheriting strategies, new individuals may adapt by copying, from known individuals, the strategies that had best results. In both cases, the frequency of each strategy in the population changes over time, and the ensemble may evolve towards a stable situation. There is also the possibility of introducing small mutations in minority, and study how they spread. Evolutionary Game Theory (EGT) provides the means to understand the why and the how of what it takes for agents with individual interests to cooperate for a common weal [45, 64].

In its simplest form, a cooperative act is metaphorically described as the act of paying a cost to convey a benefit to someone else. If two players simultaneously decide to cooperate or not, the best possible response will be to try to receive the benefit without paying the cost. In an evolutionary setting, we may also wonder why would natural selection equip selfish individuals with altruistic tendencies while it incites competition between individuals and thus apparently rewards only selfish behaviour? Several mechanisms responsible for promoting cooperative behaviour have been recently identified [65, 94]. From kin and group ties [102, 111], to different forms of reciprocity [47, 66, 68, 72, 105] and networked populations [54, 87, 89, 90, 99], several aspects have been shown to play an important role in the emergence of cooperation.

Moreover, more complex strategies based on the evaluation of interactions between third parties allow the emergence of kinds of cooperation that are immune to exploitation because then interactions are channelled to just those who cooperate. Questions of justice and trust, with their negative (punishment) and positive (help) incentives, are fundamental in games with large diversified groups of individuals gifted with intention recognition capabilities. In allowing them to choose amongst distinct behaviours based on suggestive information about the intentions of their interaction partners—these in turn influenced by the behaviour of the individual himself—individuals are also influenced by their tolerance to error or noise in the communication. One hopes that, to start with, understanding these capabilities can be transformed into mechanisms for spontaneous organization and control of swarms of autonomous robotic agents [7], these being envisaged as large populations of agents where cooperation can emerge, but not necessarily to solve a priori given goals, as in distributed AI.

With these general objectives, we have specifically studied the way players' strategies adapt in populations involved in cooperation games. We used the techniques of EGT, considered games such as the Prisoner's Dilemma and Stag Hunt, and showed how the actors participating in repeated iterations in these games can benefit from having the ability to recognize the intentions of other actors, or to establish commitments, or both, thereby leading to an evolutionary stable increase in cooperation [27, 30–32], compared to extant best strategies.

Intention recognition (IR), or abducing intent, can be implemented using Bayesian Networks (BN) [30, 77, 78], taking into account the information of current signals of intent, as well as the mutual trust and tolerance accumulated from previous one-on-one play experience—including how my previous defections may influence another's intent—but without resorting to information gathered regarding players' overall reputation in the population. A player's present intent can be understood here as how he's going to play the next round with me, whether by cooperating or defecting. Intention recognition can also be learnt from a corpus of prior interactions among game strategies [31, 32], where each strategy can be envisaged and detected as players' (possibly changing) intent to behave in a certain way [28]. In both cases, we experimented with populations with different proportions of diverse strategies in order to calculate, in particular, what is the minimum fraction of individuals capable of intention recognition for cooperation to emerge, invade, prevail, and persist. It seems to us that even basic intention recognition, and its use in the scope of cooperation and tolerance, is a foundational cornerstone where we should and indeed began at, which was naturally followed by the capacity to establish and honour commitments [33, 34], as a tool towards the successive construction of collective intentions and social organization [92, 93].

We argue that the study of these issues in minds with machines has come of age and is ripe with research opportunities, and communicate below some of the published inroads we have achieved with respect to intention recognition, to commitment and to the emergence of cooperation, involving tolerance and intolerance, in the evolutionary game theory context.

2 Intention Recognition Promotes the Emergence of Cooperation

Most studies on the evolution of cooperation, grounded on evolutionary dynamics and game theory, have neglected the important role played by a basic form of intention recognition in behavioural evolution. In this section, we address explicitly this issue, characterizing the dynamics emerging from a population of intention recognizers. We derive a Bayesian Network model for intention recognition in the context of repeated social dilemmas and evolutionary game theory, by assessing the internal dynamics of mutual trust and tolerance, accumulated from previous one-on-one play experience, between intention recognizers and their opponents, as detailed below. Intention recognizers are then able to predict the next move of their opponents based on past direct interactions, which, in turn, enables them to prevail over the most famous strategies of repeated dilemmas of cooperation, even in the presence of noise. Overall, our framework offers new insights on the complexity and beauty of behavioural evolution driven by elementary forms of cognition.

2.1 Background

Intention recognition can be found abundantly in many kinds of interactions and communications, not only in human but also many other species [101]. The knowledge about intention of others in a situation could enable to plan in advance, either to secure a successful cooperation or to deal with potential hostile behaviours [27, 29, 83, 106]. Given the advantage of knowing the intentions of others and the abundance of intention recognition among different species, it is clear that intention recognition should be taken into account when studying or modeling collective behaviour. This issue becomes even more relevant when the achievement of a goal by an individual does not depend uniquely on its own actions, but also on the decisions and actions of others, namely when individuals cooperate or have to coordinate their actions to achieve a task, especially when the possibility of communication is limited [41, 52, 107]. For instance, in population-based artificial intelligence applications [1, 7, 26], such as collective robotics and others, the inherent problem of lack of intention recognition due to the simplicity of the agents is often solved by assuming homogeneous populations, in which each agent has a perfect image of the other as a copy of their own self. Yet, the problem remains in heterogeneous agent systems where it is likely that agents speak different languages, have different designs or different levels of intelligence; hence, intention recognition may be the only way agents understand each other to secure successful cooperation or coordination among heterogeneous agents. Moreover, in more realistic settings where deceiving may offer additional profits, individuals often attempt to hide their real intentions and make others believe in pretense ones [78, 82, 88, 98, 101].

Intention recognition is defined, in general terms, as the process of becoming aware of the intention of another agent and, more technically, as the problem of inferring an agent's intention through its actions and their effects on the environment [12, 41, 51]. For the recognition task, several issues can be raised grounded on the eventual distinction between the model an agent creates about himself and the one used to describe others, often addressed in the context of the "Theory of Mind" theory, which neurologically reposes in part on "mirror neurons", at several cortical levels, as supporting evidence [46, 61, 81].

The problem of intention recognition has been paid much attention in AI, Philosophy and Psychology for several decades [8, 9, 12, 21, 51]. Whereas intention recognition has been extensively studied in small scale interactive settings, there is an absolute lack of modelling research with respect to large scale social contexts; namely the evolutionary roles and aspects of intention recognition.

2.2 Modeling Behavioural Dynamics

Our study is carried out within the framework of Evolutionary Game Theory (EGT) [45, 58]. Here, individual success (or fitness) is expressed in terms of the outcome of a 2-person game, which, in turn, is used by individuals to copy others whenever these appear to be more successful. Comparative accumulated payoffs are used to update the population: more successful individuals produce more offspring, which inherit their strategy. Equivalently, the same process can be seen as if, instead of inheriting strategies, new individuals adapt by copying strategies from acquaintances that did better. Overall, this type of dynamics can be conveniently described as an ordinary differential equation—the replicator equation [45]—, which nicely describes any simple evolutionary process.

In our work we model intention recognition within the framework of repeated interactions. In the context of direct reciprocity [47, 48, 65, 105, 108] intention recognition is being performed using the information about past *direct* interactions. We study this issue using the well-known repeated Prisoner's Dilemma (PD) [95], i.e., so that intentions can be inferred from past individual experiences. Naturally, the same principles could be extended to cope with indirect information, as in indirect reciprocity [68, 69, 72]. This eventually introduces moral judgment and concern for individual reputation, which constitutes "per se" an important area where intention recognition may play a pivotal role [35, 72]. Here, however, we shall concentrate on the simpler case of intention recognition from past experiences.

Contrary to other approaches dealing with the integration of (direct or indirect) information about the past in individual decisions, e.g. in [57, 69, 109, 110], intention recognition is performed using a Bayesian Network (BN) model. BNs have proven to be one of the most successful approaches for intention recognition [12, 21, 29, 78, 100]. Their flexibility for representing probabilistic dependencies as well as causal relations, and the efficiency of inference methods have made them an extremely powerful tool for problem solving under uncertainty [73, 74], and appropriate to deal with several probabilistic as well as causal dependencies occurring in intention recognition. We derive a Bayesian Network model for intention recognition in the context of social dilemmas, taking into account mutual trusts between the intention recognizer and his opponent. Trusts are accumulated through past interactions, assuming that intention recognizers have a memory. Greater memory sizes enable to build longer-term mutual trusts, and therefore allow better tolerance to the errors of intended actions.

The repeated (or iterated) PD is usually known as a story of tit-for-tat (TFT), which won both Axelrod's tournaments [3, 4]. *TFT* starts by cooperating, and does whatever the opponent did in the previous round. It will cooperate if the opponent cooperated, and will defect if the opponent defected. But if there are erroneous moves because of noise (i.e. an intended move is wrongly performed with a given execution error, referred here as "noise"), the performance of *TFT* declines, in two ways: (i) it cannot correct errors and (ii) a population of *TFT* players is undermined by random drift when *AllC* (always cooperate) mutants appear (which allows exploiters to grow). Tit-for-tat is then advantageously replaced by generous tit-for-tat (GTFT), a strategy that cooperates if the opponent cooperated in the previous round, but sometimes cooperates even if the opponent defected (with a fixed probability $p > 0$). *GTFT* can correct mistakes, but remains suffering the random drift; in addition, it deals with pure defectors worse than *TFT*.

Subsequently, *TFT* and *GTFT* were replaced by win-stay-lose-shift (WSLS) as the winning strategy chosen by evolution [67]. *WSLS* repeats the previous move whenever it did well, but changes otherwise. *WSLS* corrects mistakes better than *GTFT* and does not suffer random drift. However, it is exploited seriously by pure defectors.

We consider a population of constant size N. At each evolution step, a random pair of players are chosen to play with each other. The population consists of pure cooperators, pure defectors plus either of *TFT*s or of *WSLS*s or of intention recognizers who, being capable of recognizing another's intention based on the past interactions, seek the cooperators to cooperate with and to defect toward detected defectors.

Interactions are modeled as symmetric two-player games defined by the payoff matrix, used by all players. In particular, each type of player chooses to play in the same way under the same circumstances.

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

A player who chooses to cooperate (C) with someone who defects (D) receives the sucker's payoff S, whereas the defecting player gains the temptation to defect, T. Mutual cooperation (resp., defection) yields the reward R (resp., punishment P) for both players. Depending on the ordering of these four payoffs, different social dilemmas arise [56, 87, 94]. Namely, in this work we are concerned with the Prisoner's Dilemma (PD), where $T > R > P > S$. In a single round, it is always best to defect, but cooperation may be rewarded if the game is repeated. In repeated PD, it is also required that mutual cooperation is preferred over an equal probability of unilateral cooperation and defection ($2R > T + S$); otherwise alternating between cooperation and defection would lead to a higher payoff than mutual cooperation.

Before providing our intention recognition model in the framework of social dilemmas, let us provide the definition of Bayesian Network. A Bayesian Network (BN) is a pair consisting of a directed acyclic graph (DAG) whose nodes represent variables and missing edges encode conditional independencies between the variables, and an associated probability distribution satisfying the Markov assumption of conditional independence, saying that variables are independent of non-descendants given their parents in the graph [73, 74].

In a BN, associated with each node of its DAG is a specification of the distribution of its variable, say A, conditioned on its parents in the graph (denoted by $pa(A)$)—i.e., $P(A|pa(A))$ is specified. If $p(A) = \emptyset$ (A is called root node), its unconditional probability distribution, $P(A)$, is specified. These distributions are called Conditional Probability Distribution (CPD) of the BN. The joint distribution of all node values can be determined as the product of conditional probabilities of the value of each node on its parents.

In [78], a general BN model for intention recognition is presented and justified based on Heinze's intentional model [41, 100]. Basically, the BN consists of three layers: cause/reason nodes in the first layer (called *pre-intentional*), connecting to intention nodes in the second one (called *intentional*), in turn connecting to action nodes in the third (called *activity*). Intuitively, the observed actions of an agent are causally affected by his/her intentions, which are in turn causally affected by the causes/reasons for which he committed to the intentions [8, 9]. The interested readers are referred to [41, 78, 100] for detailed discussions.

Based on this general model, we present an intention recognition model in the context of the social dilemmas, taking into account the past *direct* interactions (Fig. 1). The model is described from the view of an intention recognizer (denoted by $\mathscr{I}$) with respect to a co-player (denoted by $\mathscr{J}$), whose intention (C or D) is to be recognized. A player's intentions here can be understood as the characters or types of the player: how cooperative or defective he is in general when playing with me. Saying that the co-player has intention C (resp., D) means that, in general, he intends to cooperate with me (resp., exploit or defect towards me). Thus, if he has been cooperative in the past, it is likely he will continue to cooperate in the current interaction.

$\mathscr{J}$'s intention in a given interaction is causally affected by the trust he holds towards his opponent ($\mathscr{I}$), which is accumulated over their past (observed) interactions. $\mathscr{J}$'s intention in turn has given rise to his past actions. Let $M > 0$ be the memory size of intention recognizers, i.e. they can remember their moves and their opponents' moves in the last M rounds of interaction with any specific players.

For this Bayesian Network, we need to determine the prior probability of the node *oTrust*, i.e. $P(Tr)$; the CPD table of node *Intention*—specifying the conditional probability of $\mathscr{J}$ having an intention (C or D) given the trust he holds towards his opponent ($\mathscr{I}$), i.e. $P(I|Tr)$; and the CPD table of the node *pastObs*—specifying the conditional probability of the past observations given $\mathscr{J}$'s intention (C or D), i.e. $P(O|I)$.

The accumulated payoff from all interactions (not shown here) emulates the individual *fitness* or social *success* and the most successful individuals will tend to be imitated by others, implementing a simple form of social learning [24, 94, 103].

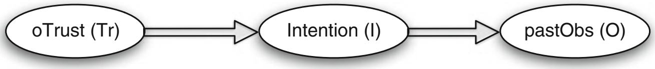

Fig. 1 Bayesian Network for Intention Recognition in Social Dilemmas. Pre-intentional level has one node, *oTrust (Tr)*, receives Boolean values, t (*true*) or f (*false*), representing the other's trust on us (the intention recognizers). Intentional level has one node, *Intention (I)*, receiving value C or D, corresponding to more cooperative and more defective, respectively, in the past. It is causally affected by *oTrust*. Activity level has one node, *pastObs (O)*, causally affected by *Intention* node. Its value is a pair (n_C, n_D) where n_C and n_D are the number of times the recognized player cooperated and defected, respectively, in the recent M (memory size) steps. *pastObs* is the only observed (evidence) node

Any player (including *IR*) can change its strategy by adopting another player's strategy with a probability defined by the Fermi distribution [64]. If a strategy has a higher (average) payoff or fitness than another, it tends to be imitated more by the other. The *IR* strategy in general has higher fitness than all others, thus it tends to by imitated by them, thereby dominating the population most of the time.

In the commonly used settings, including when interacting solely with the pure strategies (AllC and AllD) and when all considered strategies interacted with each other [47, 48, 65], IR always outperforms TFT and WSLS [30, 32]. The population spends more time in the homogeneous state of all being IRs, even in the presence of noise and of small random mutations. Furthermore, since a population of IRs is highly cooperative, it is clear that the introduction of intention recognition significantly increases the cooperation level of the population, leading to a greater social welfare.

2.3 Discussion

Using the tools of EGT, we have addressed the role played by intention recognition in the evolution of cooperation. In this work, we have shown, in a novel way, the role of intention recognition for the emergence of cooperation within the framework of the repeated Prisoner's Dilemma. Intention recognition is performed using a Bayesian Network model via computing mutual trusts between the intention recognizers and their opponents. Given the broad spectrum of problems which are addressed using this cooperative metaphor, our result indicates how intention recognition can be pivotal in social dynamics. We have shown that the intention recognition strategy prevails over the most successful existent strategies (TFT, WSLS) of the repeated PD, even when players have a very limited memory. IR deals with AllD better than TFT—the best known defector-dealer—, and corrects mistake better than WSLS—the best known mistake-corrector [65, 95]. As a result, a homogenous population of IRs has a higher level of cooperation than the ones of WSLSs and TFTs, resisting the invasion of other strategies.

In [47], it has been shown that in absence of noise, in a population of AllCs, AllDs and TFTs, the population spends most of the time in a homogeneous state of TFTs. However, as we have shown elsewhere, it is not the case if noise is present, especially under strong selection. In absence of noise, IR behaves the as well as TFT. Moreover, IRs are selected by evolution in the latter case where noise is present. We have shown that in a population of AllCs, AllDs and IRs, the population spends most of the time in homogeneous state of IRs in a broad range of scenarios and parameters, especially when the intensity of selection is strong. We have also exhibited experimentally that in a population where all the five strategies AllC, AllD, TFT, WSLS and IR are present, IRs still prevail most of the time. Therefore, together with the fact that IRs can correct mistakes better than WSLSs and TFTs, the presence of IRs would significantly increase the overall level of cooperation of the population.

Additionally, we have shown the role of a large memory size in recognizing/correcting errors, that is in recovering from ongoing persistent mutual defection that may result from a move announcement mistake, or from communication channel noise. Having a greater memory size allows to build longer-term mutual trusts/distrusts, and hence enables to better recognize erroneous moves. It then enables to better tolerate of a selfish act made by cooperative trustful individuals, and refuses to cooperate after an erroneous cooperation made by a defective untrustworthy ones. Indeed, intention recognition gives rise to an incipient mechanism of commitment formation, from which future behaviours may be assessed and trust bonds established. Overall, our work provides new insights on the complexity and beauty of behavioural evolution driven by basic, elementarily defined, forms of cognition.

3 Commitment Promotes the Emergence of Cooperation

Agents make commitments towards others, the promise to enact their play moves in a given manner, in order to influence others in a certain way, often by dismissing more profitable options. Most commitments depend on some incentive that is necessary to ensure that the action is in the agent's interest and thus, may be carried out to avoid eventual penalties [22]. The capacity for using commitment strategies effectively is so important that natural selection may have shaped specialized signaling capacities to make this possible [5, 15, 63, 82, 88, 98]. And it is believed to have an incidence on the emergence of morality [85]. Assuming cooperation to be, at best, just the result of individuals' purely competitive strategies can make it conceptually unstable [71], most especially in non-iterated or history-free interactions. And it seems possible that the spread of simplistic notions, rooted in science, about the evolutionary origins of social relationships could foster a trend to make these relationships more conflicted, and society more brutal. An antidote is an evolutionary approach to behaviour that incorporates a capacity for mutual commitment, shown advantageous for all concerned [63], even in non-iterated or memory-free settings.

Our goal is to examine, through EGT [45, 58, 94], how the most simple of commitment strategies work, and how they can give rise to the emergence of cooperation. We shall do so in the setting of the non-iterated Prisoner's Dilemma (PD).

In a nutshell, convincing others of one's credibility in a commitment proposal amounts to submit to options that change the incentives of the situation. These options, namely commitment cost and penalty for defaulting, can be expressed by the payoffs specified in a game. When opponent players observe meticulously such payoffs, and realize that compliance with a proposed commitment is in the proposing player's best interests, then, given any opponent player's open option to commit, these may change their expectations and behaviour accordingly, and adopt as a result a strategy which either accepts commitment proposals or ignores them. In general, there are four main reasons to believe a commitment will be fulfilled [63]: (i) a commitment can be self-reinforcing if it is secured by incentives intrinsic to the situation; (ii) a commitment can be secured by external incentives controlled by third parties; (iii) a commitment can be backed by a pledge of reputation; and (iv) a commitment can be reinforced by internal emotional motives. The first two types are secured in much the same way a loan is secured by a collateral. They objectively change the situation so that fulfillment becomes in the individual's best interests. The latter two types do not change the objective contingencies; they are subjective commitments in that they may involve a continued option of reneging, according to some or other stance extraneous to the game's given payoff matrix.

In this section, we provide a new EGT model showing that individuals tend to engage in commitments, which leads to the emergence of cooperation even without assuming repeated interactions. The model is characterized by two key parameters: the punishment cost of failing commitment imposed on either side of a commitment, and the cost of managing the commitment deal. Our analytical results and extensive computer simulations show that cooperation can emerge if the punishment cost is large enough compared to the management cost.

3.1 Model

In our EGT setting, the game's payoff matrix summarily ingrains and expresses in its structure the impingement of all such contingencies [33, 34]. For instance, often a capacity for commitment allows individuals to act in ways that reap the benefits of image scoring through maintaining a reputation, or the access of others to a social history of prior interactions. In this study, for simplicity but also for exhibiting the purity and power of the commitment mechanism, we ignore the effect of repeated interactions [104, 105], and of any reputation [68, 72] associated with particular individuals. We have shown [33, 34] that the simplest of core commitment mechanisms can improve cooperation, and leave any other complications for the future, most promisingly how commitment can be combined with and reinforce other known mechanisms of cooperation, for instance, intention recognition [30–32]. And perhaps surprisingly we can do so. Thus, no credibility of commitment is taken into account [11] beyond that which is expressed in a game's payoff matrix. No reputation appraisal of the commitment proposer is made by its co-player, and no historical or social data is even available to do so. Each pairwise interaction is purely based on fixed individual strategies that might involve commitment or the lack thereof. Also, no cheater or deceit detection or intention recognition is in place [30, 31, 49]. Nevertheless, systematic unconditional bluffing on the part of a player is a possible fixed feature of its strategy, in the sense that, from the start, the player does not intend to fulfill commitments. In our commitment model players defaulting on their commitments, be they the proposing or the accepting party, are subject to evolutionary disadvantage for a wide range of parameters. Commitments come at a price: players must pay to propose commitment, but commitment acceptors that default are penalized a compensation value in favor of the proposer. We have shown, with the model below, that more elaborate commitment strategies are not strictly necessary for commitment to become evolutionarily advantageous. Neither an aptitude for higher cognition, nor for empathy, nor for mind reading are needed. These aptitudes would only be required for more sophisticated forms of commitment, scaffolded atop the core one. We have explained the evolution, in a population, of the capacity for a simple form of commitment as the result of otherwise being excluded from a group of committed promise abiding cooperators, in the sense that this strategy tends to invade the game playing population under rather general conditions.

Let us consider a commitment variant of the Prisoner's Dilemma game in which a new type of cooperator (denoted by COM_C) that, before each interaction, asks the co-player whether it commits to cooperate. If the co-player does not so commit, there is no interaction. Both players get 0. Otherwise, if the co-player commits, they then go on to play with each other in the present interaction. If the co-player keeps to its commitment, both players obtain the reward payoff, R. Otherwise (if the co-player fails its commitment), the proposing or focal player obtains the sucker payoff, S, and its co-player obtains the temptation payoff, T. However, the one that fails the commitment will suffer a penalty cost, and its non-defaulting co-player gains a compensation for the potential loss due to its default of fulfilling the commitment. For simplicity, we assume that these two amounts (penalty and compensation) are equal, being denoted by δ. The penalty cost can be a real monetary one, e.g., in the form of prior debit (e.g., in the case of accommodation rental) or of a subsequent punishment cost (e.g., commitment was performed in terms of a legal contract, and one who fails commitment must pay a cost to compensate for the other), or an imaginary abstract value, e.g., public spread of good/bad reputation (bad reputation for the one that fails, and sympathy for the other), or even an emotional suffering [22, 43, 63, 85]. How this cost is set up depends on the types of commitment at work, or the reason for which the commitment is believed to be fulfilled (see beginning of Sect. 3), which topic is beyond the scope of this paper. However, various techniques can be seen in [43, 91].

Two players that defect in an interaction obtain the punishment payoff, P. For setting up a commitment, the proposer must pay a small management cost, ε. The cost of proposing and setting up the commitment might be high, but it is reasonable to assume that this cost is quite small compared to the mutual benefit of a cooperation strategy guaranteeing commitment, $\varepsilon < < R$.

We consider a finite population of a constant size, consisting of four strategies: COM_C (as described above), C (always cooperates, without proposing to commit), D (always defects, and does not commit when being asked to), and D_COM (always defects, though commits when being asked to). Here, for the sake of exposition, we assume that cooperators, including COM_C and C players, always commit whenever being asked to since they are better off to do so, as cooperation is their default choice, and reasonable commitment deals only are proposed.

In each round, two random players are chosen from the population for an interaction. For the row player, the (average) payoff matrix is consequently rendered as:

$$\begin{array}{c c} & \begin{array}{cccc} COMC & C & D & DCOM \end{array} \\ \begin{array}{c} COMC \\ C \\ D \\ DCOM \end{array} & \begin{pmatrix} R-\varepsilon/2 & R-\varepsilon & -\varepsilon & S+\delta-\varepsilon \\ R & R & S & S \\ 0 & T & P & P \\ T-\delta & T & P & P \end{pmatrix} \end{array} \tag{1}$$

Note that when a COM_C interacts with another COM_C, only one of them pays the cost of having proposed commitment, ε (e.g., the arbitrary one that proposes). Therefore, the average payoff of a COM_C in playing with another COM_C is, $R - \varepsilon/2$.

All in all, our study exhibits that, in spite of the absence of repeated interactions, reputation effect, network reciprocity, as well as group and kin selection, the strategy of commitment proposal may enable the emergence of cooperation, even under the presence of noise. By imposing a high cost for failing a commitment, when compared to the cost of setting up or managing the commitment deal, the commitment cooperative agents COM_C can get rid of the fake committers (D_COM) as well as avoid being exploited by the pure defectors (D), while playing approximately equally well against the pure cooperators (C). The results of this study suggest that our specialized capacity for commitment, which might have been shaped by natural selection [63], consists in a capacity for managing to impose a high cost of punishment, whether it is monetary or of abstract emotional or reputation value, with a relatively small cost.

Furthermore, the analytical results, supported by extensive computer simulations, showing the explicit relationships between the factors involved in the commitment mechanism would clearly provide important insight into the design of multi-agent systems resorting to commitments to facilitate cooperation [11, 13, 38, 43, 52, 91, 112, 113].

3.2 Related Work

Evolutionary explanations of commitment, particularly its role in the evolution of cooperation, have been actively sought for and discussed in several fields, including Psychology and Philosophy [5, 11, 15, 22, 43, 63, 85]. But there are only a few computational models that show the evolutionary advantages of commitment in problems where cooperative acts are beneficial [82, 88, 98]. In addition, often models rely on repeated interactions or long-term relationships [5, 15], alike the conditions where Triver's direct reciprocity [105] may play a role. Here we provide an analytic model in the framework of evolutionary game theory showing that, with the availability of the mechanism of commitment, cooperation can emerge even without assuming repeated interactions, or the availability of player reputation.

We note that there is a significant difference between our commitment model and works by others on costly punishment [17, 18, 40, 70, 80]. A commitment deal must be agreed by both sides of it in advance, thereby giving credibility and justification to punish any defaulting player. In addition, the prior agreement gives rise to compensation—the amount of which, in some cases, is agreed explicitly in advance—to the non-defaulting player. This compensation for the non-defaulting player is the significant difference that makes successful those players using the commitment strategy, while those using the costly punishment strategy have only a narrow margin of efficiency [70]; does not stand out as a winning strategy [17]; nor does it promote cooperation at all when taking into account antisocial punishment [42, 80]. The compensation might bring benefit to the commitment strategists once an appropriate deal would be arranged. This suggests that although costly punishment, whether it is social or antisocial, might not promote the evolution of cooperation, what we call 'justified' punishment, which is warranted by an appropriate commitment deal, does, so that bluffing committers are in the limit scourged. This kind of punishment might not be costly at all, and can even bring net benefit to its upholder, hence leading to the emergence of cooperation.

Last but not least, it is undoubtedly important to mention the extensive literature of AI and multi-agent systems research on commitment, e.g., [11, 13, 38, 43, 52, 91, 112, 113]. The main concern therein is how to formalize different aspects of commitment and how a commitment mechanism can be implemented in multi-agent interactions to enhance them (e.g. for improved collaborative problem solving [113]), especially in the context of game theory. In contradistinction, our concern is in the nature of an evolutionary explanation of commitment, particularly how it can promote the emergence of cooperation. More importantly, our evolutionary study of the commitment mechanism leads to insights about the global influence of the mechanism within a (large) population of agents, thereby enabling improvement for the design of multi-agent systems operating upon commitments [13, 112, 113].

3.3 Discussion

Within the general game theory concept of commitment, or intention manifestation, several distinctions can help separate different subtypes. In particular, some commitments are upfront promises of a next move that can help, while others are upfront threats of a subsequent move that can harm. Commitments can be conditional or unconditional. Threats are usually attempts to influence another person's next move by stating a conditional subsequent move, and that's how we may envisage them. Promises are more likely to be unconditional, and that's how we may conceive of them, though more generally they can be conditional on the other fulfilling a matching promise.

Commitments can also be just towards oneself, taking into account the evolution of possible futures afforded by actions and events, and the individual's prior and post preferences, in what might be classically seen as a game against nature.

In [75, 76], three different types of individual commitment—hard, revocable, and momentary—are studied in such an evolution context. Let us recall that commitment, in the context of game theory, is a device or mechanism to decide the outcome with the other party [91]. Schelling distinguishes between commitment pure and simple and commitment that takes the form of a threat. What he calls "ordinary" commitment corresponds, in game theory, to the making of an opening announcement in a sequential play, which we dub preemptive, just before both players make their actual move. To constitute a preemption, a player's announcement action must be irrevocable, that is a promise that is assuredly kept. Preemptive commitment is not necessarily profitable, because it hinges on the opponent's actual move. Schelling however does not assume the other type of commitment as a "threat", which pertains to a player's move in reaction to the opponent's move. Threats, being conditional, may be of the "if-then-else" form, and can thus combine a threat and a promise, the latter albeit implicit whenever there are just two possible moves. We prefer instead to label "reactive" such so-called threat commitments. In the game context, these occur when the player with the last move irrevocably pledges to respond, in a specified but contingent way, to the opponent's prior choice [44].

In a nutshell, some players can be "preemptive" committers—those that always propose and always accept proposed commitments—, others may be "reactive" committers—those that always make a "reactive" statement and comply with the implicit requests in such statements—, while other players, though accepting to commit nevertheless default on their commitment, and even others simply omit and ignore preemptive or reactive commitments in their strategies—they might for instance be persistent defectors or persistent cooperators as we have seen, or, for that matter, follow any other strategy ignorant of commitment. Moreover, in iterated games, commitments can concern future rounds and not just the present one.

We purport to have shown that a simple commitment abiding cooperative strategy can be evolutionarily advantageous even in a non-iterated game setting. But much remains to be explored. In the more general setting and to avoid confusion, it can be helpful to distinguish, even if only conceptually, between "execution moves" and "pre-play moves" [44]. The terms first move and last move then always refer exclusively to execution moves—the choices that actually generate the payoffs. In contrast, commitments come earlier with respect to execution moves: they are pre-play moves. A preemptive commitment is a pre-play move that allows the player making it to take the first execution move. A reactive commitment, although also a pre-play move, can be made only by the player who has the last execution move. In either case, by giving up on his or her choice through committing, the commitment player leaves the opponent with "the last clear chance to decide the outcome" [91].

In our present game setting, however, there was no need to make the distinction between the first and the second to play, because each possible player strategy move is exhibited and fixed from the start, as expressed and ingrained in the payoff matrix. By so introducing the several committed unconditional move strategies—though the payoff is of course conditional on the opponent's move—, we can emulate what would happen in a round if a move sequence actually existed. Put briefly, our commitment model is of the simplest kind and, moreover, it is brought to bear solely on the very next move fold of a pair of players, with no history available on prior commitments. Nevertheless, it captures core features of commitment, namely the high cost of defaulting to discourage false commitment, and thus make it plausible, and a comparatively small but non-zero cost of commitment proposal to lend it initial credibility. On top of this core model more elaborate models affording commitment can subsequently be rooted, including those involving delayed deceit.

What's more, commitment (or intention manifestation) and intention recognition, are but two sides of a coin really, and their future joint study in the EGT setting is all but unavoidable [27]. It has become increasingly obvious that maximizing reproductive success often requires keeping promises and fulfilling threats, even when that requires in turn sacrifices regarding individual short-term interests. That natural selection has shaped special mental capacities to make this possible seems likely, including a capacity for commitment [63] and for intention recognition [30, 31]. The commitment stance goes yet further, and many aspects of human groups seem shaped by effects of commitments and intention recognition, namely group boundaries, initiation rituals, ideologies, and signals of loyalty to the group [96–98]. Conversely, many aspects of groups seem to exist largely to facilitate commitment to cooperate and to limit the utility of coercive threats.

The generalized ability for commitment to support cooperative interaction is an important aspect of plasticity in human behaviour, and humans support their dealmaking in lots of ways. The law is full of instances of people using techniques of commitment to establish the honesty of their intentions, namely through a variety of contracts [23]. Institutions themselves are supported on committal contracts, and the law of the land proffers methods for constituting and of accountability of social institutions [93].

Given our rigorous approach's inroad results, we believe they lend promise to that further studies of commitment will benefit greatly from rigorous models that allow for their analytical study and computer simulation, and in particular within the fold of EGT for the better to examine the emergence of complex social behaviour.

4 Intention Recognition, Commitment, and Evolution of Cooperation

Individuals make commitments towards others in order to influence others to behave in certain ways. Most commitments may depend on some incentive that is required to ensure that the action is in the agent's best interest and thus, should be carried out to avoid eventual penalties. Similarly, individuals may ground their decision on an accurate assessment of the intentions of others. Hence, both commitments and intention recognition go side by side in behavioural evolution. Here, we analyze the role played by the co-evolution of intention recognition plus the emergence of commitments, in the framework of the evolution of cooperative behaviour. We resort to tools of evolutionary game theory in finite populations, showing how the combination of these two aspects of human behaviour can enhance the emergent fraction of cooperative acts under a broad spectrum of configurations.

There are cases where it is difficult, if not impossible, to recognize the intentions of another agent. It might be your first interaction with someone in your life, and you have no information about him/her which can be used for intention recognition. You also might know someone well, but you still might have very little relevant information in a given situation to predict the intentions with high enough confidence. Furthermore, you might also have abundance of relevant observations about him/her, but he/she is so unpredictable that you have rarely managed to predict his/her true intention in the past. In all such cases, the strategy of proposing commitment, or intention manifestation, can help to impose or clarify the intentions of others. Note that *intention is choice with commitment* [8, 14, 84]. Once an agent intends to do something, it must settle on some state of affairs for which to aim, because of its resource limitation and in order to coordinate its future actions. Deciding what to do established a form of commitment [14, 84]. Proposing a commitment deal to another agent consists in asking it to express or clarify its intentions.

One of the commitments we all know is marriage. By giving up the option to leave someone else, spouses gain security and an opportunity for a much deeper relationship that would be impossible otherwise [20, 63], as it might be risky to assume a partner's intention of staying faithful without the commitment of marriage. Though suggestive, this simplistic view of marriage also reveals some of the simplifications of the model. A marriage is indeed a commitment between partners. However, it is also a signal to the social group of the partners? cohesion with the group, and a signal that each partner sends to himself or herself, validating the choice of staying in the relationship.

A contract is another popular kind of commitment, e.g. for an apartment lease [20]. When it is risky to assume another agent's intention of being cooperative, arranging an appropriate contract provides incentives for cooperation. However, for example in accommodation rental, a contract is not necessary when the cooperative intention is of high certainty, e.g. when the business affair is between close friends or relatives.

Having said this, arranging a commitment deal can be useful to encourage cooperation whenever intention recognition is difficult, or cannot be performed with sufficiently high certainty. On the other hand, arranging commitments is not free, and requires a specific capacity to set it up within a reasonable cost (for the agent to actually benefit from it) [62, 63]—therefore it should be avoided when opportune. In the case of marriage, partners sometimes choose to stay together without an official commitment when it might be too costly (e.g., it could be against parents' or families' wish, or it may need to be in secret because of their jobs) and/or they strongly trust each other's faithfulness (e.g., because of emotional attachment [19, 20]). In short, a combination of the two strategies, those of commitment and of intention recognition, seems unavoidable. Nevertheless, intention recognition without actual commitment can be enhanced by costly engagement gifts, in support of sexual selection and attachment [39, 60]. Furthermore, social emotions can act as ersatz commitment [19].

Here, we start from the model [33] of commitment formation (described in the previous section), characterized by two key parameters: a punishment cost of failing commitment imposed on either side of a commitment deal, and the cost of managing it. On top of that model, again using EGT, we show that combining intention recognition and commitment strategies in a reasonable way can lead to the emergence of improved cooperation, not able to be achieved solely by either strategy. Our study seeks what is a reasonable combination of commitment and intention recognition.

We shall do so in the setting of the Prisoner's Dilemma (PD). It will be seen from our model that, in most of the cases, there is a wide range of combination of the intention recognition and commitment strategies, which leads to a strategy that performs better than either strategy solely—in the sense that the population spends more time in the homogeneous state of agents using that strategy [40, 47]. Our results suggest that, if one can recognize intentions of others with high enough confidence or certainty, one should rely more on it, especially when it is difficult to reach to a conceivably strong commitment deal. It helps to avoid the unnecessary cost of arranging and managing the deal. That is, in a transparent world where people have nothing to hide from each other, contracts are unnecessary.

On the other hand, when intention recognition with high precision is difficult (due to, e.g. environment noise, agents have great incentives to hide intentions, or there is not enough observed actions), one should rely more on the commitment strategy, particularly if a reasonable deal can be envisaged.

4.1 A Minimal Model Combining Intention Recognition and Commitment

We provide a new strategy, IRCOM, which combines the two strategies, those of intention recognition and commitment. In an interaction, IRCOM recognizes the intention (cooperates or defects) of its co-player [30]. A confidence level, *cl*, is assigned to the recognition result. It defines the degree of confidence (here in terms of probability) that IRCOM predicts the co-player's intention correctly.

Note that in AI the problem of intention recognition has been paid much attention for several decades, and the main stream is that of probabilistic approaches [2, 6, 10, 12, 41]. They tackle the problem by assigning probabilities to conceivable intentions (conditional on the current observations), based on which the intentions are ranked. Similarly to [2, 6, 28], in our model, a degree of confidence, *cl*, in terms of a probability measure, is assigned to intentions.

In general, *cl* follows some probability distribution. As in a real intention recognition problem, the distribution should depend on the intention recognition method at work (how efficient it is), the environment IRCOM lives in (is it supportive for gathering relevant information for the recognition process, e.g. observability of co-players' direct and indirect interactions, perception noise, population structure), etc. For example, we can consider different distributions satisfying that the longer IRCOM survives, the more precisely or confidently it performs intention recognition; or, considering the repeated interaction setting in the framework of the iterated PD, the more IRCOM interacts with its co-player, the better it can recognize the co-player's intention (see intention recognition models for the iterated PD in [30–32]).

We model *cl* by a continuous random variable X with probability density function $f(x, U)$, where U is a vector characterizing the factors that might influence *cl*, including the efficiency of the intention recognition model at work, the environmental factors (e.g., noise, population structure), and the interaction setting (repeated, one-shot, etc.).

If IRCOM is confident enough about the intention recognition process and result, that is *cl* is greater than a given, so-called, *confidence threshold* $\theta \in [0, 1]$, then in the current interaction IRCOM cooperates if the recognized intention of the co-player is to cooperate, and defects otherwise. The prediction is wrong with probability $(1 - cl)$. For simplicity, we assume that the prediction is a (continuous) random variable, Y, uniformly distributed in $[0, 1]$. Hence, the probability that IRCOM utilizes intention recognition, but with an incorrect and correct prediction, respectively, can be written as joint probability distributions [25, 36].

If $cl \leq \theta$, i.e. IRCOM is not confident enough about its intention prediction, it behaves the same as COM_C (see above). The greater θ is, the more cautious IRCOM is about its intention recognition result. Obviously, if $\theta = 1$, IRCOM behaves identically to COM_C ; and if $\theta = 0$, IRCOM behaves identically to a (pure) intention recognizer [30, 31].

We now replace COM_C with IRCOM, considering a population of four strategies, IRCOM, C, D, and D_COM. For the row player, the (average) payoff matrix reads $M = \theta M_1 + M_2$, where M_2 is the payoff matrix when IRCOM utilizes the intention recognition strategy, i.e. in the case $cl > \theta$. To derive M_2, we consider the case that *cl* has a uniform distribution in the interval $[0, 1]$, i.e. $f(x, U) = 1$ for $x \in [0, 1]$ and 0 otherwise.

The main subject of our published analysis is to address, given the payoff entries of the PD, and the parameters of the commitment deal IRCOM can manage, how confident about the intention recognition result IRCOM should be in order to make a decision, without relying on the commitment proposing strategy. That is, if there is an optimal value of θ for an IRCOM to gain greatest net benefit.

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 [30–32], where concrete intention recognition models are deployed.

However, when a quite strong commitment deal can be envisaged, arranging it can still glean some evolutionary advantage. But in case only weak commitment deals can be arranged, it is then more beneficial to rely, even exclusively, on the intention recognition strategy, should it be efficient enough.

4.2 Discussion

A general implication of our analysis is that an appropriate combination of the two strategies of commitment and intention recognition often leads to a strategy that performs better than either one solely. It is advantageous to rely on the intention recognition strategy (when reaching sufficiently high confidence about its result) because it helps to avoid the cost of arranging and managing commitment deals, especially when no strong deals can be arranged or envisaged.

This result has a similar implication to that obtained in [50], where the authors show that overconfidence might give evolutionary advantage to its holders. In our model, an IRCOM can gain extra net benefit if it is a little overconfident (that is, when using sufficiently small θ), taking risk to rely on intention recognition result instead of arranging some commitment deal. Differently, because in our model IRCOM is further guaranteed by an efficient strategy of commitment, being over-overconfident (that is, using too small θ) and relying exclusively on intention recognition might prevent it from opportunely gaining benefit from the commitment strategy—especially in case the intention recognition model at work is not efficient. It said, the performance of overconfident individuals [50] can be enhanced by relying on the commitment strategy when they need to muster overly high courage (say, in order to decide to claim some resource).

In the framework where intention recognition is difficult and of high risk, for example, climate change negotiation [59, 79, 86], military setting—comprising a lot of bluffing [53, 91]—and international relationships [55], our model suggests arranging a strong commitment deal.

4.3 Conclusions

Assume simply that we are given an intention recognition method, that affords us a degree of confidence distribution *cl* about its predictions, with regard to the intentions of others, and hence their future actions, typically on the basis of their seen actions and surrounding historical and present circumstances. Assume too some commitment model is given us about providing mutual assurances, and involving an initial cost and a penalty for defaulting.

We have shown how to combine together one such general intention recognition method, with a specific commitment model defined for playing the Prisoner's Dilemma (PD), in the setting of Evolutionary Game Theory (EGT), by means of a single payoff matrix extended with a new kind of player, IRCOM, which chooses whether to go by the result of its intention recognition method about a co-player's next move, or to play by the commitment strategy, depending on whether its level of confidence on the intention prediction *cl* exceeds or not some a given confidence threshold θ. Our results indicate that IRCOM is selected by evolution for a broad range of parameters and confidence thresholds.

Then we have studied, for a variety of *cl* and θ, in the context of PD in EGT, how IRCOM performs in the presence of other well-known non-committing strategies (always cooperate, C, and always defect, D) – plus the strategy that commits when being asked to, but always defects, D_COM. Analytical and simulation results show under which circumstances, for different *cl* and θ, and distinct management and punishment costs, ε and δ, does the new combined strategy IRCOM prove advantageous and to what degree. And does indeed, IRCOM proves to be adaptably advantageous over those other just mentioned strategies and in all circumstances from a quite small confidence level onwards.

Much remains to be done with respect to further consideration of combining the two strategies of intention recognition and commitment. The two go often together, and not just in the basic way we have examined. Actually they are two sides of one same coin, one side being an attempt to identify an intention, the other being the manifestation of an intention. For one, we only considered the case where intention recognition comes first in order to decide on a commitment proposal. But, in general, once a commitment is made, intention recognition is a paramount method to follow up on whether the commitment will be honoured, on the basis of detecting or otherwise not the intermediate actions leading up to commitment fulfillment. Furthermore, the information about commitments can be used to enhance intention recognition itself.

It seems to us that intention recognition, and its use in the scope of commitment, is a foundational cornerstone where we should begin at, naturally followed by the capacity to establish and honour commitments, as a tool towards the successive construction of collective intentions and social organization [92, 93]. Finally, one hopes that understanding these capabilities can be useful in the design of efficient self-organized and distributed engineering applications [7], from bio- and socio-inspired computational algorithms, to swarms of autonomous robotic agents.

5 Coda

Evolutionary Psychology and Evolutionary Game Theory provide a theoretical and experimental framework for the study of social exchanges.

Recognition of someone's intentions, which may include imagining the recognition others have of our own intentions, and may comprise not just some error tolerance, but also a penalty for unfulfilled commitment, can lead to evolutionary stable win/win equilibriums within groups of individuals, and perhaps amongst groups. The recognition and the manifestation of intentions, plus the assumption of commitment—even whilst paying a cost for putting it in place—, are all facilitators in that respect, each of them singly and, above all, in collusion.

What is more, by means of joint objectives under commitment, one might promote the inclusion of heretofore separate groups into more global ones. The overcoming of intolerance shall benefit from both levels of manifest interaction—individual and group-wise.

We have argued that the study of these issues, of minds as evolving machines, has come of age and is ripe with research opportunities—including epistemological—and have communicated in some detail here some of the inroads we have explored, and have pointed to the much more detailed published results of what we have achieved, with respect to intention recognition, commitment, and mutual tolerance, within the overarching evolutionary game theory context.

The work of many other authors has also been emphasized and been given references, so the interested reader may easily begin to delve into this fascinating area, and follow up on its very active ongoing exploration and applications potential.

Acknowledgments This article is in great part a summary and cursory overview of the authors' detailed joint work mentioned in the publications. TAH acknowledges support from FCT-Portugal (grant SFRH/BD/62373/2009) and the FWO Vlaanderen (Postdoctoral fellowship). FCS acknowledges support from FCT-Portugal (PTDC/FIS/101248/2008, PTDC/MAT/122897/2010 and INESC-ID's PEst-OE/EEI/LA0021/2011). We thank all three reviewers for their very helpful indications.

References

1. Ampatzis, C., Tuci, E., Trianni, V., Dorigo, M.: Evolution of signaling in a multi-robot system: categorization and communication. Adapt. Behav. **16**(1), 5–26 (2008)
2. Armentano, M.G., Amandi, A.: Goal recognition with variable-order markov models. In: Proceedings of the 21st International Joint Conference on, Artificial Intelligence, pp. 1635–1640 (2009)
3. Axelrod, R.: The Evolution of Cooperation. Basic Books, New York (1984). ISBN 0-465-02122-2
4. Axelrod, R.: The evolution of cooperation. Science **211**, 1390–1396 (1981)
5. Back, Istvan, Flache, Andreas: The adaptive rationality of interpersonal commitment. Ration. Soc. **20**(1), 65–83 (2008)
6. Blaylock, N., Allen, J.: Statistical goal parameter recognition. In: Zilberstein S., Koehler J., Koenig S. (eds.) Proceedings of the 14th International Conference on Automated Planning and Scheduling (ICAPS'04), pp. 297–304. AAAI (2004)
7. Bonabeau, E., Dorigo, M., Theraulaz, G.: Swarm Intelligence: From Natural to Artificial Systems. Oxford University Press, USA (1999)
8. Bratman, M.E.: Intention, Plans, and Practical Reason. The David Hume Series, CSLI (1987)
9. Bratman, M.E.: Faces of Intention: Selected Essays on Intention and Agency. Cambridge University Press (1999)

10. Bui, H., Venkatesh, S., West, G.: Policy recognition in the abstract hidden markov model. J. Artif. Intell. Res. **17**, 451–499 (2002)
11. Castelfranchi, C., Falcone, R.: Trust Theory: A Socio-Cognitive and Computational Model (Wiley Series in Agent Technology). Wiley (2010)
12. Charniak, E., Goldman, R.P.: A Bayesian model of plan recognition. Artif. Intell. **64**(1), 53–79 (1993)
13. Chopra, A.K., Singh, M.P.: Multiagent commitment alignment. In: Proceedings of the 8th International Joint Conference on Autonomous Agents and MultiAgent Systems (AAMAS '09), pp. 937–944 (2009)
14. Cohen, P.R., Levesque, H.J.: Intention is choice with commitment. Artif. Intell. **42**(2–3), 213–261 (1990)
15. de Vos, H., Smaniotto, R.: Reciprocal altruism under conditions of partner selection. Ration. Soc. **13**(2), 139–183 (2001)
16. Deacon, T.W.: The hierarchic logic of emergence: Untangling the interdependence of evolution and self-organization. In: Weber, H.W., Depew, D.J. (eds.) Evolution and Learning: The Baldwin Effect Reconsidered. MIT Press, Cambridge, MA (2003)
17. Dreber, A., Rand, D.G., Fudenberg, D., Nowak, M.A.: Winners don't punish. Nature **452**(7185), 348–351 (2008)
18. Fehr, E., Gachter, S.: Altruistic punishment in humans. Nature **415**, 137–140 (2002)
19. Frank, R.H.: Passions Within Reason: The Strategic Role of the Emotions. W. W. Norton and Company, New York (1988)
20. Frank, Robert H.: Cooperation through emotional commitment. In: Nesse, R.M. (ed.) Evolution and the Capacity for Commitment, pp. 55–76. Russell Sage, New York (2001)
21. Geib, C.W., Goldman, R.P.: A probabilistic plan recognition algorithm based on plan tree grammars. Artif. Intell. **173**(2009), 1101–1132 (2009)
22. Gintis, H.: Beyond selfishness in modeling human behavior. In: Nesse, Randolf M. (ed.) Evolution and the Capacity for Commitment. Russell Sage, New York (2001)
23. Goodenough, O.R.: Law and the biology of commitment. In: Nesse, R.M. (ed.) Evolution and the Capacity for Commitment, pp. 262–291. Russell Sage, New York (2001)
24. Szabó, G., Tőke, C.: Evolutionary prisoner's dilemma game on a square lattice. Phys. Rev. E **58**, 69–73 (1998)
25. Gut, A.: An Intermediate Course in Probability, 2nd edn. Springer Publishing Company, Incorporated, New York (2009)
26. Gutierrez, A., Campo, A., Santos, F.C., Monasterio-Huelin, F., Dorigo, M.: Social odometry: imitation based odometry in collective robotics. I. J. Adv. Robot. Syst. **2**(6), 129–136 (2009)
27. Han, T.A.: Intention recognition, commitments and their roles in the evolution of cooperation. Ph.D. thesis, Department of Informatics, Faculty of Sciences and Technology, Universidade Nova de Lisboa (May 2012)
28. Han, T.A., Pereira, L.M.: Context-dependent incremental intention recognition through Bayesian network model construction. In: Nicholson, A. (ed.) Proceedings of the Eighth UAI Bayesian Modeling Applications Workshop (UAI-AW 2011), vol. 818, pp. 50–58. CEUR Workshop Proceedings (2011)
29. Han, T.A., Pereira, L.M.: Intention-based decision making via intention recognition and its applications. In: Guesgen, H., Marsland, S. (eds.) Human Behavior Recognition Technologies: Intelligent Applications for Monitoring and Security. IGI Global, (forthcoming) (2013)
30. Han, T.A., Pereira, L.M., Santos, F.C.: Intention recognition promotes the emergence of cooperation. Adapt. Behav. **19**(3), 264–279 (2011)
31. 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), pp. 1684–1689. AAAI (2011)
32. Han, T.A., Pereira, L.M., Santos, F.C.: Corpus-based intention recognition in cooperation dilemmas. Artif. Life j. **18**(4), 365–383 (2012)

33. Han, T.A., Pereira, L.M., Santos, F.C.: The emergence of commitments and cooperation. In: Proceedings of the 11th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'2012), pp. 559–566. ACM (2012)
34. Han, T.A., Pereira, L.M., Santos, F.C.: Intention recognition, commitment, and the evolution of cooperation. In: Proceedings of IEEE Congress on Evolutionary Computation, pp. 1–8. IEEE Press, June 2012
35. Han, T.A., Saptawijaya, A., Pereira, L.M.: Moral reasoning under uncertainty. In: Proceedings of the 18th International Conference on Logic for Programming, Artificial Intelligence and Reasoning (LPAR-18), pp. 212–227. Springer LNAI 7180 (2012)
36. Han, T.A., Traulsen, A., Gokhale, C.S.: On equilibrium properties of evolutionary multiplayer games with random payoff matrices. Theor. Popul. Biol. **81**(4), 264–272 (June 2012)
37. Hardin, G.: The tragedy of the commons. Science **162**, 1243–1248 (1968)
38. Harrenstein, P., Brandt, F., Fischer, F.: Commitment and extortion. In: Proceedings of the 6th International Joint Conference on Autonomous Agents and MultiAgent Systems, AAMAS '07, ACM, New York, USA (2007)
39. Haselton, M.G., Buss, D.M.: Error management theory: a new perspective on biases in cross-sex mind reading. J. Pers. Soc. Psychol. **78**(1), 81–91 (2001)
40. 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)
41. Heinze, C.: Modeling intention recognition for intelligent agent systems. Ph.D. thesis, The University of Melbourne, Australia (2003)
42. Herrmann, Benedikt, Thöni, Christian, Gächter, Simon: Antisocial Punishment Across Societies. Science **319**(5868), 1362–1367 (2008)
43. Hirshleifer, J.: Game-theoretic interpretations of commitment. In: Nesse, Randolf M. (ed.) Evolution and the Capacity for Commitment, pp. 77–93. Russell Sage, New York (2001)
44. Hirshleiffer, J.: There are many evolutionary pathways to cooperation. J. Bioecon. **1**(1), 73–93 (1999)
45. Hofbauer, J., Sigmund, K.: Evolutionary Games and Population Dynamics. Cambridge University Press (1998)
46. Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J.C., Rizzolatti, G.: Understanding emotions in others: mirror neuron dysfunction in children with autism spectrum disorders. PLoS Biology **3**(3), e79 (2005)
47. Imhof, L.A., Fudenberg, D., Nowak, M.A.: Evolutionary cycles of cooperation and defection. Proc. Nat. Acad. Sci. U S A **102**, 10797–10800 (2005)
48. Imhof, L.A., Fudenberg, D., Nowak, M.A.: Tit-for-tat or win-stay, lose-shift? J. Theor. Biol. **247**(3), 574–580 (2007)
49. Janssen, M.: Evolution of cooperation in a one-shot prisoner?s dilemma based on recognition of trustworthy and untrustworthy agents. J. Econ. Behav. Organ. **65**(3–4), 458–471 (2008)
50. Johnson, D.D.P., Fowler, J.H.: The evolution of overconfidence. Nature **477**(7364), 317–320 (2011)
51. Kautz, H., Allen, J.F.: Generalized plan recognition. In: Proceedings of the Conference of the American Association of Artificial Intelligence (AAAI'1986), pp. 32–38. AAAI (1986)
52. Kraus, S.: Negotiation and cooperation in multi-agent environments. Artif. Intell. **94**(1–2), 79–98 (1997)
53. Leeds, Brett A.: Alliance reliability in times of war: explaining state decisions to violate treaties. Int. Organ. **57**(04), 801–827 (2003)
54. Lindgren, K., Nordahl, M.G.: Evolutionary dynamics of spatial games. Physica D: Nonlinear Phenom. **75**(1–3), 292–309 (1994)
55. Lockhart, Charles: Flexibility and commitment in international conflicts. Int. Stud. Quart. **22**(4), 545–568 (1978)
56. Macy, M.W., Flache, A.: Learning dynamics in social dilemmas. Proc. Nat. Acad. Sci. U S A **99**, 7229–7236 (2002)

57. Masuda, Naoki, Ohtsuki, Hisashi: A theoretical analysis of temporal difference learning in the iterated prisoner's dilemma game. Bull. Math. Biol. **71**, 1818–1850 (2009)
58. Maynard-Smith, J.: Evolution and the Theory of Games. Cambridge University Press, Cambridge (1982)
59. Milinski, M., Semmann, D., Krambeck, H.J., Marotzke, J.: Stabilizing the Earth's climate is not a losing game: supporting evidence from public goods experiments. Proc. Nat. Acad. Sci. U S A **103**, 3994–3998 (2006)
60. Miller, Geoffrey F., Todd, Peter M.: Mate choice turns cognitive. Trends Cogn. Sci. **2**(5), 190–198 (1998)
61. Nakahara, K., Miyashita, Y.: Understanding intentions: through the looking glass. Science **308**(5722), 644–645 (2005)
62. Nesse, R.M.: Evolution and the Capacity for Commitment. Russell Sage Foundation series on trust, Russell Sage (2001)
63. Nesse, Randolf M.: Natural selection and the capacity for subjective commitment. In: Nesse, Randolf M. (ed.) Evolution and the Capacity for Commitment, pp. 1–44. Russell Sage, New York (2001)
64. Nowak, M.A.: Evolutionary Dynamics: Exploring the Equations of Life. Harvard University Press, Cambridge, MA (2006)
65. Nowak, M.A.: Five rules for the evolution of cooperation. Science **314**(5805), 1560 (2006). doi:10.1126/science.1133755
66. Nowak, M.A., Sigmund, K.: Tit for tat in heterogeneous populations. Nature **355**, 250–253 (1992)
67. 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)
68. Nowak, M.A., Sigmund, K.: Evolution of indirect reciprocity. Nature **437**,1291–1298 (2005)
69. Ohtsuki, H., Iwasa, Y.: The leading eight: social norms that can maintain cooperation by indirect reciprocity. J. Theor. Biol. **239**(4), 435–444 (2006)
70. Ohtsuki, H., Iwasa, Y., Nowak, M.A.: Indirect reciprocity provides only a narrow margin of efficiency for costly punishment. Nature **457**(7601), 79–82 (2009)
71. Oyama, S.: Evolution's Eye: A Systems View of the Biology-Culture Divide. Duke University Press, Durham (2000)
72. Pacheco, J.M., Santos, F.C., Chalub, F.A.C.C.: Stern-judging: a simple, successful norm which promotes cooperation under indirect reciprocity. PLoS Comput. Biol. **2**(12), e178 (2006)
73. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, San Mateo (1988)
74. Pearl, J.: Causality: Models, Reasoning, and Inference. Cambridge University Press (2000)
75. Pereira, L.M., Han, T.A.: Evolution prospection. In: Proceedings of International Symposium on Intelligent Decision Technologies (KES-IDT'09), pp. 51–63. Springer Studies in Computational Intelligence 199, 2009
76. Pereira, L.M., Han, T.A.: Evolution prospection in decision making. Intell. Decis. Technol. **3**(3), 157–171 (2009)
77. Pereira, L.M., Han, T.A.: Intention recognition via causal Bayes networks plus plan generation. In: Progress in Artificial Intelligence, Proceedings of 14th Portuguese International Conference on Artificial Intelligence (EPIA'09), pages 138–149. Springer LNAI 5816, Oct 2009
78. Pereira, L.M., Han, T.A.: Intention recognition with evolution prospection and causal Bayesian networks. In: Computational Intelligence for Engineering Systems 3: Emergent Applications, pp. 1–33. Springer (2011)
79. Nichola, R., Aitken, D.: Uncertainty, rationality and cooperation in the context of climate change. Climatic Change **108**(1), 47–55 (2011)
80. Rand, D.G., Nowak, M.A.: The evolution of antisocial punishment in optional public goods games. Nature Commun. **2**:434 (2011)

81. Rizzolatti, G., Craighero, L.: The mirror-neuron system. Annu. Rev. Neurosci. **27**, 169–192 (2004)
82. Robson, A.: Efficiency in evolutionary games: Darwin, Nash, and the secret handshake. J. Theor. Biol. **144**(3), 379–396 (1990)
83. Roy, O.: Intentions and interactive transformations of decision problems. Synthese **169**(2), 335–349 (2009)
84. Roy, O.: Thinking before acting: intentions, logic, rational choice. Ph.D. thesis, ILLC Dissertation Series DS-2008-03, Amsterdam (2009)
85. Ruse, M.: Morality and commitment. In: Nesse, Randolf M. (ed.) Evolution and the Capacity for Commitment, pp. 221–236. Russell Sage, New York (2001)
86. Santos, F.C., Pacheco, J.M.: Risk of collective failure provides an escape from the tragedy of the commons. Proc. Nat. Acad. Sci. U S A **108**(26), 10421–10425 (2011)
87. Santos, F.C., Pacheco, J.M., Lenaerts, T.: Evolutionary dynamics of social dilemmas in structured heterogeneous populations. Proc. Nat. Acad. Sci. U S A **103**, 3490–3494 (2006)
88. Santos, F.C., Pacheco, J.M., Skyrms, B.: Co-evolution of pre-play signaling and cooperation. J. Theor. Biol. **274**(1), 30–35 (2011)
89. Santos, F.C., Pinheiro, F.L., Lenaerts, T., Pacheco, J.M.: The role of diversity in the evolution of cooperation. J. Theor. Biol. **299**, 88–96 (2012)
90. Santos, F.C., Santos, M.D., Pacheco, J.M.: Social diversity promotes the emergence of cooperation in public goods games. Nature **454**, 214–216 (2008)
91. Schelling, T.C.: The Strategy of Conflict. Oxford University Press, London (1990)
92. Searle, J.R.: The Construction of Social Reality. The Free Press, New York (1995)
93. Searle, J.R.: Making the Social World: The Structure of Human Civilization. Oxford University Press (2010)
94. Sigmund, K.: The Calculus of Selfishness. Princeton University Press (2010)
95. Sigmund, K., De Silva, H., Traulsen, A., Hauert, C.: Social learning promotes institutions for governing the commons. Nature **466**, 7308 (2010)
96. Skyrms, B.: Evolution of the Social Contract. Cambridge University Press (1996)
97. Skyrms, B.: The Stag Hunt and the Evolution of Social Structure. Cambridge University Press (2003)
98. Skyrms, B.: Signals: Evolution, Learning, and Information. Oxford University Press (2010)
99. Szabó, G., Fáth, G.: Evolutionary games on graphs. Phys. Rep. **446**(4–6), 97–216 (2007)
100. Tahboub, K.A.: Intelligent human-machine interaction based on dynamic Bayesian networks probabilistic intention recognition. J. Intell. Robot. Syst. **45**, 31–52 (January 2006)
101. Tomasello, M.: Origins of Human Communication. MIT Press, Cambridge (2008)
102. Traulsen, A., Nowak, M.A.: Evolution of cooperation by multilevel selection. Proc. Nat. Acad. Sci. U S A **103**(29), 10952 (2006)
103. Traulsen, A., Nowak, M.A., Pacheco, J.M.: Stochastic dynamics of invasion and fixation. Phys. Rev. E **74**, 11909 (2006)
104. Trivers, R.: Deceit and Self-Deception: Fooling Yourself the Better to Fool Others. Penguin Books, Limited (2011)
105. Trivers, R.L.: The evolution of reciprocal altruism. Q. Rev. Biol. **46**, 35–57 (1971)
106. van Hees, M., Roy, O.: Intentions and plans in decision and game theory. In: Verbeek, B. (ed.) Reasons and Intentions, pp. 207–226. Ashgate Publishers, Aldershot (2008)
107. Van Segbroeck, S., de Jong, S., Nowé, A., Santos, F.C., Lenaerts, T.: Learning to coordinate in complex networks. Adapt. Behav. **18**, 416–427 (2010)
108. Van Segbroeck, S., Pacheco, J.M., Lenaerts, T., Santos, F.C.: Emergence of fairness in repeated group interactions. Phys. Rev. Lett. **108**, 158104 (2012)
109. Vukov, J., Santos, F.C., Pacheco, J.M.: Incipient cognition solves the spatial reciprocity conundrum of cooperation. PLoS ONE **6**(3), e17939 (March 2011)
110. Wang, S., Szalay, M.S., Zhang, C., Csermely, P.: Learning and innovative elements of strategy adoption rules expand cooperative network topologies. PLoS ONE **3**(4), e1917, 04 2008

111. West, S.A., Griffin, A.A., Gardner, A.: Evolutionary explanations for cooperation. Curr. Biol. **17**, R661–R672 (2007)
112. Winikoff, M.: Implementing commitment-based interactions. In: Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS '07, pp. 868–875. ACM, New York, USA (2007)
113. Wooldridge, M., Jennings, N.R.: The cooperative problem-solving process. J. Logic Comput. **9**, 403–417 (1999)