

## The implications of shared identity on indirect reciprocity

Wafi Bedewi, Roger M. Whitaker, Gualtiero B. Colombo, Stuart M. Allen & Yarrow Dunham

To cite this article: Wafi Bedewi, Roger M. Whitaker, Gualtiero B. Colombo, Stuart M. Allen & Yarrow Dunham (2020): The implications of shared identity on indirect reciprocity, Journal of Information and Telecommunication, DOI: [10.1080/24751839.2020.1741858](https://doi.org/10.1080/24751839.2020.1741858)

To link to this article: <https://doi.org/10.1080/24751839.2020.1741858>



© 2020 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



Published online: 19 Mar 2020.



Submit your article to this journal [↗](#)



Article views: 257






View related articles [↗](#)



View Crossmark data [↗](#)



# The implications of shared identity on indirect reciprocity\*

Wafi Bedewi <sup>a</sup>, Roger M. Whitaker <sup>a</sup>, Gualtiero B. Colombo<sup>a</sup>, Stuart M. Allen <sup>a</sup> and Yarrow Dunham<sup>b</sup>

<sup>a</sup>School of Computer Science & Informatics, Cardiff University, Cardiff, UK; <sup>b</sup>Department of Psychology, Yale University, New Haven, CT, USA

## ABSTRACT

The ability to sustain indirect reciprocity is an example of collective intelligence. It is increasingly relevant to future technology and autonomous machines that need to function in a coalition. Indirect reciprocity involves providing benefit to others without guaranteeing a future return. The identity through which an agent presents itself to others is fundamental, as this is how the reputation of an agent is considered. In this paper, we examine the sharing of identity between agents, which is an important and frequently overlooked issue when considering indirect reciprocity. We model an agent's identity using traits, which can be shared with other agents, and offer a basis for an agent to change their identity. Through this approach, we determine how shared identity affects cooperation, and the conditions through which cooperation can be sustained. This also helps us to understand how and why behavioural strategies involving identity function are put in place, such as whitewashing. The framework offers the opportunity to assess the interplay between the sharing of traits and the cost, in terms of reduced cooperation and opportunities for shirkers to benefit.

## ARTICLE HISTORY

Received 15 January 2020  
Accepted 10 March 2020

## KEYWORDS

Identity; cooperation; indirect reciprocity; reputation; traits

## 1. Introduction

Cooperation is a sophisticated form of collective intelligence where individuals become incentivised to help one another and benefit from a coalition. One particularly interesting but challenging form of cooperation is *indirect reciprocity*, which is complex because it involves donating to a third party without any guarantee of future reciprocation. Cooperation in this form involves a small cost to the donor, and a much larger benefit to the recipient. This is a hallmark of human behaviour that leads to a societal benefit, by providing a resource through which unrelated individuals support each other (Alexander, 1987; Bear & Rand, 2016).

**CONTACT** Wafi Bedewi  [bedewiwa@cardiff.ac.uk](mailto:bedewiwa@cardiff.ac.uk)

\*This paper is an extension of work that was presented in ICCCI 2019 (Bedewi W., Whitaker R.M., Colombo G.B., Allen S.M., Dunham Y. (2019) Modelling stereotyping in cooperation systems. In: Nguyen N., Chbeir R., Exposito E., Anioré P., Tra-wiński B. (eds) *Computational collective intelligence. ICCCI 2019*. Lecture Notes in Computer Science, vol 11683. Springer, Cham) which has been selected to be extended and published in a fast track for the 'Journal of Information and Telecommunication' (JTIT) published by Taylor & Francis. This paper extends our previous work by focusing on identity, and allowing the model to explore the copying of traits along with strategies by individuals.

© 2020 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Extensive research has been successful in establishing conditions and mechanisms that promote indirect reciprocity. However, as machines are developed that feature cognition and autonomy, interest in cooperation is reaching beyond humans (de Melo et al., 2019). Transportation is just one emerging example where technology, through autonomous vehicles, will encounter cooperative decision making (Imbsweiler et al., 2018). This scenario features latent indirect reciprocity, such as when one driver allows another to manoeuvre in traffic. Journeys in congestion often depend on this, such as when exiting a T-junction, without which safe progress would be impossible in many cities.

Beyond technological scenarios, persistent human scenarios such as inter-group conflict (Tajfel & Turner, 1979) continue to motivate the exploration of cooperation, and the basis for it being sustained. The decision on whether or not to cooperate, when called upon, is the fundamental issue. The personal *identity* is central to this – in human terms personal identity represents the self-image that defines an individual and that is projected to others. This can be a function of traits and characteristics that are relevant to the individual, including the different groups to which the individual belongs. Human group identity can be either formed around immutable traits such as family, nationality, race, ethnicity, religion, or more temporal ones as a supported sport team, or even based on the current working environment, friendship and familiarity links (Tajfel, 1974).

This opens up the possibility of components of identity being shared and may lead to ‘group mind’ behaviours like in-group favouritism and out-group bias (Swann & Buhrmester, 2015). It is from identity that *reputation* is derived. Reputation provides a currency through which cooperation can be recognized and signalled (Nowak & Sigmund, 2005), allowing individuals to leverage future help when needed (Molleman et al., 2013). Human group identity can become an important component of the extent to which the reputation of an individual is formed and recognized independently from the personal actions. In extreme situations, this can lead to the loss of any personal identity, where reputation is fully merged with that of the group(s), leading to social phenomena such as stereotyping (Hales, 1998). In recent times reputation systems have also emerged to support decision making in diverse areas of e-commerce. In auction systems, a seller’s reputation proves fundamental in the willingness of buyers to decide whether or not to place a bid (Melnik & Alm, 2002). Furthermore, e-commerce reputation also serves beyond auction settings to signal the quality of product and services (Resnick et al., 2000). This information has significant value as a ‘public good’ (Wasko & Faraj, 2000). There are several other areas of work in multi-agent systems where the focus is to engineer protocols or rules that seek to ensure cooperation is followed. These approaches aim to disincentivize deviation from behaviours that benefit the public good (Wu et al., 2016).

The origins of reputation systems come from behaviour in groups with humans being adept at using reputation to assess the integrity of others (Suzuki & Akiyama, 2005), as a means to promote their survival. This allows groups to function, and humans are skilled at creating heuristics, or cognitive short cuts, that allow them to find potential cooperators without extensive deliberation. However, these cognitive short cuts can also have negative implications. In the context of driving dynamics, for example, the type of vehicle, its manufacturer, the age, gender or other characteristics of the driver may well influence whether one driver helps another. While this may appear insignificant, in the wider human context this behaviour can have a major impact, being responsible for bias that fuels stereotyping (Galinsky & Moskowitz, 2000), resulting in potentially unwarranted discrimination and the spread of prejudice (Oakes &

Turner, 1980). Divisive social consequences may result (Kawakami et al., 2017), leading to categorization, where the reputation that an individual incurs has no alignment to their actual behaviour. This is a key component in theories concerning inter-group conflict. These issues are also transferred to technological scenarios, depending on the capacity of machines to align with human bias or foster it themselves (Whitaker et al., 2018).

### 1.1. Contribution

This work contributes to understanding how the sharing of identity impacts on cooperation. This is achieved by modelling shared traits, that carry reputation in their own right from which an individual's reputation is derived based on the extent that particular traits represent their identity. The approach involves agent-based simulation, where agents have some freedom in how they adapt their behaviour based on probabilistically copying the strategy and possibly the traits of others, based on perceived success. This approach allows us to explore conditions that either promote or impede cooperation. It should not be confused with agent-based approaches in knowledge engineering, where protocols are sought that allow cooperation to be enforced based on individual behaviour (e.g. Wu et al., 2016). It can be noted that the vast majority of social-psychological treatments related to identity and stereotyping assume a single unique trait per individual, despite increasing demands to capture the ground truth of social organization (Bowleg, 2017).

Indirect reciprocity is the basis for our investigation, but other forms of cooperation could also be applied. The approach is novel because models of indirect reciprocity conventionally assume that each individual is represented by a unique reputation: in other words, an individual's behaviour is entirely identified and judged by their own actions. Our model goes beyond this one-to-one mapping, allowing reputations to be implicitly shared by different actors. In the context of cooperation, this means that individuals become dependent on the donation behaviour of others for an element of their reputation. Furthermore, our framework does not assume that 'groups' to which individuals belong are mutually exclusive. Reputations are calculated on traits, any number of which can be held by an individual. This better represents the fluidity that is seen in the real world, where individuals are rarely totally defined by a single group affiliation and group identity, but may be represented as a combination of characteristics and affiliations.

We examine how both repeated sharing of the same trait and across multiple traits affects the emergence of cooperation. This represents a general scenario where traits are fixed and persistent (i.e. the agent cannot change traits). In contrast, we also examine how the ability to change traits and pursue the traits of those deemed most successful, affects cooperation. This is an evolutionary form of 'whitewashing', where identity becomes a strategic component that is mutable in pursuit of payoff. These results help us to understand how the structure of traits and the freedom of agents in changing them affects cooperation through indirect reciprocity. This extends our work on stereotyping and contributions to previously limited insights in this direction.

## 2. Key related literature

This research focuses on indirect reciprocity, groups and the role of reputation. Indirect reciprocity is frequently considered in the context of the donation game, where an

agent has to make a decision on whether or not to provide a donation. This results in a cost  $c$  to the donor, and a benefit  $b$  to the recipient, and necessarily  $c < b$  (Brandt et al., 2007; Nowak & Sigmund, 2005). Reputation systems act to signal an agent's overall donation behaviour to the wider population. Because other agents may use an agent's reputation in deciding when or not to donate, there is an incentive for all potential recipients to maintain reputation at a sufficient level to yield future donations (Fehr, 2004; Milinski et al., 2002; Wedekind & Milinski, 2000).

Critical within reputation systems are assessment rules. These are the criteria by which a donor's reputation is adjusted in light of their actions, and therefore govern the extent of reward over penalty. In this sense, they have been considered as a model for morality (Alexander, 1987). Three main alternatives for assessment of cooperative action are *image scoring*, *standing* and *judging*. Sugden (1986) first developed standing, which was originally conceived for binary reputations. This assessment rule effectively classifies each individual in the population as either good or bad, penalizing the good if they donate to the bad.

Image scoring (Nowak & Sigmund, 1998; Wedekind & Milinski, 2000) presented the first significant alternative, where reputation is simply incremented or decremented in response to donation or defection respectively. A limitation of image scoring is that those who choose not to cooperate with defectors may be unfairly labelled as less cooperative (Leimar & Hammerstein, 2001; Panchanathan & Boyd, 2003). Consequently, with their roots in the work of (Sugden, 1986), *standing* (Panchanathan & Boyd, 2003) and *judging* (Brandt & Sigmund, 2004) have emerged as the alternatives that capture 'legitimate shirking' (Fishman, 2003; Nowak & Sigmund, 2005; Rand & Nowak, 2013). These discrimination rules have mainly been studied assuming that reputation has a binary representation (Brandt et al., 2007; Ohtsuki & Iwasa, 2006), although this was generalized for standing in Whitaker et al. (2016).

The overwhelming convention is that individuals hold their own individual reputation with similarity of reputation only introduced to address uncertainty (e.g. Leimar & Hammerstein, 2001; Nowak & Sigmund, 1998). Deviation from this has occurred in the biological literature, specifically concerning the plausibility of group selection such as in Wilson (1975) and Smith (1964). These models assume that individuals belong to precisely one group, and it is the group entity that determines whether or not individuals propagate to future generations. This was largely dismissed by the biological literature but was revisited when the idea of multi-level selection was proposed by Wilson and Sober (1994), where individual and group identity coexist and may promote cooperation (Nowak et al., 2010). Reputation systems can feature in this context, allowing individuals to potentially switch between individual and group reputations (Masuda, 2012; Suzuki & Akiyama, 2005). However, this remains a controversial theory, as discussed by Pinker (2012).

Psychological processes of categorization are well seen in human behaviour, and work relating to groups and cooperation has featured consideration of both in-group bias (Fu et al., 2012; Hammond & Axelrod, 2006) and out-group prejudice (Brewer, 1999; Whitaker et al., 2018), while not necessarily invoking the use of a group reputation. These contributions reflect the disposition of individuals to differentiate, either implicitly or explicitly, based on their strong identification with self-similar individuals (Launay & Dunbar, 2015). Stereotyping is a related extension of this, where third party individuals

are categorized together through a perception of common identity (Galinsky & Moskowitz, 2000). This is well known to be a divisive phenomenon in the human world (Dovidio et al., 1998; Tajfel et al., 1971; Turner et al., 1987).

In the case of reputation systems, only a few contributions consider categorization. In Baranski et al. (2006) the impact of group reputation is considered through multi-agents. Here, the concept of group reputation is shared by all individuals within a group when they interact with out-group members. This is calculated as the average of all individual reputations in a group and assumes that group reputation is an aggregation of the behaviour of individuals. Similarly in Masuda (2012), a group structure is proposed where individuals interact within their groups using a personal reputation. When they play out-group, individuals adopt a group-level reputation. This model also assumes that reputation is binary. These models do not allow for individuals to share subsets of traits, or aspects of their identity, and depend on individuals belonging to a single group. Our approach is to allow individuals to have a more complex composition of their identity, based on the assessment of multiple traits against which reputations are maintained.

Trait and set membership have also received attention as simple signalling mechanisms to promote the evolution of cooperation. Without the use of reputation, these elements have been regarded as abstract tags that are sufficient to incentivise some level of cooperation, which is known as the green beard effect (Nowak, 2006; Riolo et al., 2001). The evolution of set theory shows that more complex set structures can promote the emergence of cooperation even in absence of other incentives (Hamilton, 1964; Lieberman et al., 2005; Nowak & May, 1992). Tarnita et al. (2009) proposed a model based on the evolution of sets where the degree of shared membership is based on the overlapping of sets of multiple traits. In this model, the interaction is limited to traits that they have in common with others. An individual's strategy and set membership updates under evolutionary settings. However, individuals only have one strategy which is to cooperate or defect. Moderate levels of cooperation can be sustained with a limited mutation on traits (Nathanson et al., 2009; Tarnita et al., 2011). Similarly, Li et al. (2016) adopted the same model to study evolutionary dynamics of minimum-effort coordination games in structured populations.

In Gao et al. (2018), the authors describe a model in which individuals are in groups and interactions may occur between in-group members and across groups. The model does not allow for individuals to have membership in more than one group. Individuals have two strategies that enable them to act differently towards in-group and out-group members. Although their model does not rely on reputation, it allows for mutation during the reproduction phase. Mutation, in this case, occurs on traits and strategies. Their model is adapted from a simplified prisoner's dilemma.

The option to change identity leads to opportunities for agents to gain an advantage. *Whitewashing* is a term that has been used to describe the action of agents who change their identities in order to avoid punishment from other agents (Feldman & Chuang, 2005a). The term has been mostly used to describe this action within peer-to-peer reputation systems where users have been able to replace their pseudonyms to escape from any punishment due to their bad reputation. Whitewashing or re-entry attacks enable free-riders to restore their reputation to gain some short-term payoff, (Hoffman et al., 2009). Only limited research has studied the subject within an evolutionary perspective to gather an understanding of whitewashing in cooperative situations (Feldman & Chuang, 2005b).

Whitewashing reduces the opportunity for agents to accumulate a bad reputation and opens up opportunities for defection as a consequence.

### 3. Model

The simulation model that is introduced pays attention to the structure of reputation that agents hold when engaged in a cooperative dilemma (indirect reciprocity). Rather than individuals holding their own unique reputation, or being identified by a single group membership, the concept of *traits* is used to represent how individuals may be perceived as belonging to groups and present a personal identity as a consequence. Traits are features that are held by agents and represent identifiable characteristics. All agents have at least one trait, and each trait may belong to one or more agents. Unless otherwise specified by the experiment, the traits are assumed to be immutable.

Rather than reputation being associated with individual agents or mutually exclusive groups, it is assumed that each trait  $t \in T$  has associated with it a reputation  $r_t$ , and an agent  $i$  derives its personal reputation  $r^i$  from the reputations of the traits associated with  $i$ . Specifically, let  $T_i$  (with  $|T_i| > 0$ ) denote the associated set of traits for agent  $i$ , and  $r^i$  its reputation:

$$r^i = \sum_{t \in T_i} r_t / |T_i|$$

In other words, an agent's reputation is the average of the reputation of its associated traits. Consequently, any individual element of reputation relies exclusively on an agent being the only one to hold a certain trait.

This arrangement allows identity to be considered: traits belonging to an agent and shared by others are components of personal identity and are used as a proxy for their individual reputation. Furthermore, traits do not necessarily partition agents into mutually exclusive sets or groups, providing a useful generalization. This approach is applied using cooperation in the form of indirect reciprocity.

#### 3.1. Indirect reciprocity

The donation game is adopted, which is a subclass of the mutual aid game (Sugden, 1986) where the donor incurs a cost with no guarantee of reciprocation from the beneficiary, or any other individual. This is modelled through prosocial donations which result in a cost  $c$  to the donor agent and a benefit  $b$  to the recipient, where  $b > c > 0$ . There are wide-ranging models for indirect reciprocity (e.g. Leimar & Hammerstein, 2001; Nowak & Sigmund, 1998; Ohtsuki & Iwasa, 2006; Takahashi & Mashima, 2006), however, this work uses the recent and remarkably simple approach of *social comparison* of reputation proposed in Whitaker et al. (2016). This follows the human disposition to make relative judgments about the standing of others.

Each agent  $i$  carries a binary vector of variables  $(s_i, u_i, d_i)$  which represents  $i$ 's current *action rule* with respect to  $i$ 's donation behaviour when it is called upon to consider making a donation to another agent  $j$ . The action rule indicates whether or not  $i$  donates when similarity ( $s_i$ ), upward ( $u_i$ ), or downward self-comparison ( $d_i$ ) is observed by  $i$  in respect of  $j$ 's reputation ( $r^j$ ), as compared to  $i$ 's own reputation value ( $r^i$ ). Similarity



in self-comparison is identified when  $r^j = r^i$ , upward self-comparison occurs when  $r^j > r^i$ , and downward self-comparison occurs when  $r^j < r^i$ .

Periodically each agent updates its action rule through social learning, as a consequence of observing others in the population. Similarly, each agent may update its trait to reflect this observation. It is known (Whitaker et al., 2016) that evolution promotes the action rule (1, 1, 0), allowing agents to discriminate against those having a lower reputation than themselves, thereby representing a relative threat.

### 3.2. Updating reputation

Every time an agent  $i$  is called to play the donation game with a potential recipient  $j$ ,  $i$ 's donation decision depends on the agent's action rule, and the reputation of traits associated with are updated as a consequence of the outcome. The concept of standing is used. If  $i$  donates, then  $r_t$  is incremented, for all  $t \in T_i$ . If  $r^j \geq r^i$  and  $i$  defects then the reputation of trait  $t$ ,  $r_t$  is decremented, for all  $t \in T_i$ . This means that an individual's actions equally affect the traits by which it is represented. Note that the updating approach ensures that a reduction in reputations does not occur when  $i$  fails to donate and  $j$  is of a lesser reputation, providing a defense against shirkers. Each trait's reputation is capped and allowed to vary in the integer range  $[-5, 5]$ .

### 3.3. Performing the game

The donation game is performed on a set of agents  $A$  representing a population of individuals, in this case  $|A| = 100$ . Each agent  $i$  has four key fundamental attributes: its set of traits  $T_i$ , its action rule  $(s_i, u_i, d_i)$ , its reputation  $r^i$  and its fitness  $f_i$ . Fitness represents the economic payoff as the accumulation of costs and benefits that are paid and received by  $i$  over the current generation. A generation involves making 5000 random selections of a potential recipient  $j$ , from the population, to play the donation game. Let  $N_j = \{x \in A - \{j\} : T_j \cap T_x \neq \emptyset\}$  be the set of agents that share at least one trait with agent  $j$  and  $\bar{N}_j = \{x \in A - \{j\} : T_j \cap T_x = \emptyset\}$ . The potential donor  $i$  is selected at random from the set  $N_j$  with probability  $s$  and from the set  $\bar{N}_j$  with probability  $1-s$ . If no suitable donors are found then  $i$  is randomly selected from  $A - \{j\}$ .

For an agent  $j$ , the potential donor agent  $i$  is selected from the sub-population having at least one trait from  $T_j$ , with probability  $s$ . Here  $s$  is a global parameter (not to be confused with  $s_i$ ) that governs the extent to which an agent is disposed to playing in-group (i.e. with similar others).

At the end of a generation, reproduction occurs. This can be thought of as social learning where agents probabilistically copy the action rules of others, taking into account the success of other agents based on their fitness. Specifically, each agent  $i$  in the population copies the action rule of another agent  $j$  randomly, weighted by  $f_j / \sum_{k=1}^n f_k$  (i.e. roulette wheel), upon which  $i$  adopts  $j$ 's action rule for the next generation.

At this point, mutation is applied to each element of an action rule with probability  $1/100$ . Prior to commencing a new generation, fitness  $f_i$  is set to zero ( $f_i = 0, \forall i$ ) and for all traits  $t$ ,  $r_t = 0$  is set. Throughout a  $c/b$  ratio of 0.7 is applied. 100,000 generations are performed and the simulation is principally evaluated by comparing the total number of instances of cooperation (i.e.  $i$  donating to  $j$  in a donation game)



across all generations. Average figures of cooperation over five randomly seeded runs are used.

To explore the effects of freedom in changing identity, we also perform experiments where agents are additionally able to copy traits, as well as the action rule, of others, probabilistically based on payoff. This allows an agents identity to evolve, influenced by the success of others. Similarly, in this scenario, mutation on each trait is selected with a probability of 1/100 after each cycle of reproduction. The pseudo code is summarized in Algorithm 1.

---

**Algorithm 1** Performing the game
 

---

```

Set of Agents  $A$ ; Set of traits  $T$ ; cost  $C$ ; benefit  $B$ 
Generate initial population of agents
Assign set of traits  $T_i$  to each agent  $i$ , as defined by the experiment
Assign  $(s_i, u_i, d_i)$  randomly from the eight possible instances
# Perform evolutionary simulation
for number of generations  $M = 100000$  do
  Set  $f_i = 0 \ \forall i \in A$  and  $r_t = 0 \ \forall t \in T$ 
  for number of iterations  $m = 5000$  do
    # Selection
    Select recipient  $j \in A$  at random;
    Let  $p \leftarrow U(0,1)$ 
    if  $p < s \mid N_j > 0$  then
      Select donor  $i$  at random from  $N_j$ 
    else if  $p \geq s \mid \bar{N}_j > 0$  then
      Select donor  $i$  at random from  $\bar{N}_j$ 
    else
      Select donor  $i$  at random from  $A - \{j\}$ 
    end if
    # Action Rules
    if  $(r^j = r^i \text{ and } s_j = 1)$  or  $(r^j > r^i \text{ and } u_j = 1)$  or  $(r^j < r^i \text{ and } d_j = 1)$  then
       $i$  donates
       $r_t \leftarrow \min(5, r_t + 1)$ ;
       $f_i \leftarrow f_i - C$ ;  $f_j \leftarrow f_j + B$ 
    else
       $i$  defects
      if  $r^j \geq r^i$  then
         $r_t \leftarrow \max(-5, r_t - 1)$ 
      end if
    end if
  end for
  # Reproduction stage
  Generate new population proportionally to the individual payoff
  Apply mutation to each agent in the new population
end for

```

---

## 4. Experiments

The model allows different types of reputation sharing with other agents based on the trait ( $s$ ) that are held in common. We say that an agent is *dependent* if it shares at least one trait with another agent. Otherwise, the agent is *independent*. Furthermore, if an agent  $i$  is such that  $|T_i| > 1$  then  $i$  is a *multi-trait* agent. Otherwise,  $i$  is a *single-trait* agent.

We consider three ways in which the structure of shared identity can be composed. In Section 4.1 we consider the effect of dependent single-trait agents on the evolution of cooperation, assuming that each agent's traits remain fixed throughout. In Section 4.2 we consider the effect of a dependent multi-trait agent on the evolution of cooperation,

again assuming that each agent's traits remain fixed throughout. Finally, in Section 4.3, we allow agents to probabilistically change their traits at the reproduction stage, based on payoff.

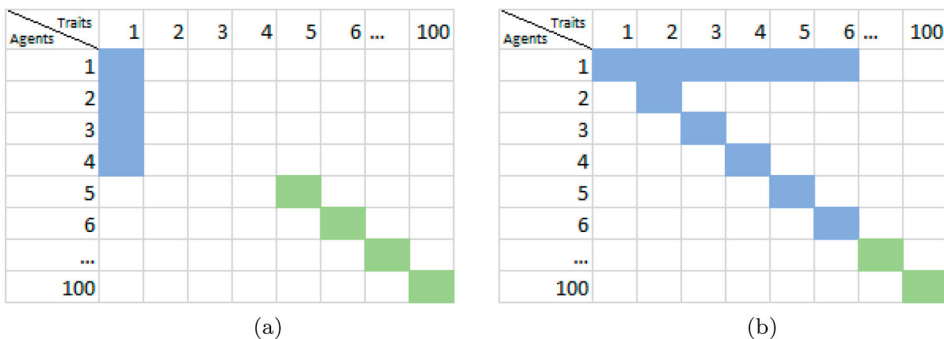
#### 4.1. Dependent single-trait identities

We consider the effects of a single common trait being shared by a set of single-trait agents. Let  $G_1$  be the set of all agents  $i$  having  $T_i = \{1\}$ . An example schema for this arrangement of traits is shown in Figure 1(a). We experiment to determine the maximum size of  $G_1$  through which cooperation can be sustained. Note that if all agents are single-trait and independent, their reputation is based entirely on their own past interactions and the results in Whitaker et al. (2016) are replicated. At the other extreme, if all agents are dependent and share a single trait, then agents are (almost) entirely judged on the actions of others, and a greater incentive to defect can be expected. The results of increasing the size of  $G_1$  is shown in Figure 2(a), alongside varying  $s$ , the probability that agent's play with those having at least one trait in common.

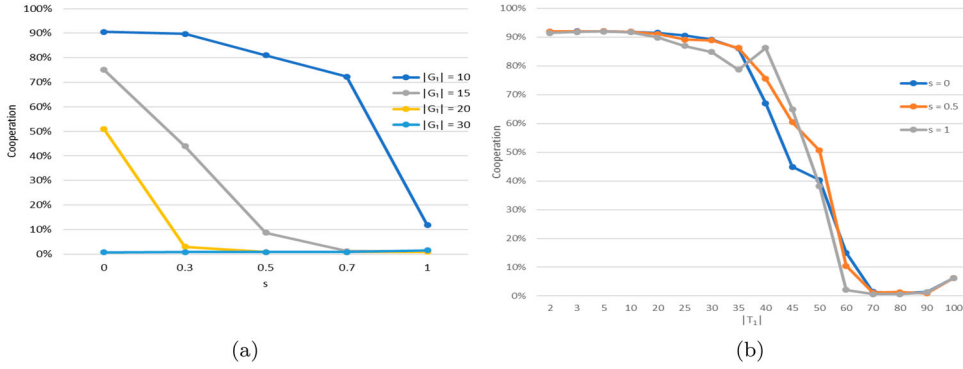
Two patterns emerge: firstly cooperation declines rapidly when at least 15 dependent single-trait agents share a common trait. Secondly, the average cooperation declines as  $s$  increases.

Dependent single-trait agents lack a distinguishable personal reputation, which means that the reputational benefit of donation is shared with others while the cost is borne by the individual. This stereotyping effect provides an opportunity for defective strategies to take hold, where free riders can benefit from enjoying a shared reputation without donating. However, this cannot be sustained at scale, leading to the global collapse of cooperation. As the reputation of a shared trait increases, there is greater opportunity for exploitation by free riders.

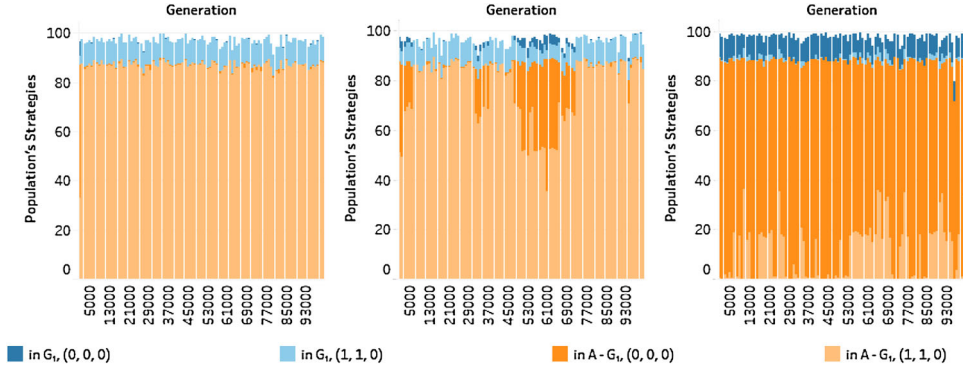
Figure 3 shows the action rules (defection strategy (0, 0, 0) and discrimination strategy (1, 1, 0)) that occur across the populations considered as subsequent generations occur, with  $|G_1| = 10$ . This is presented for three different values of  $s$  (0, 0.5, 1). The discrimination strategy dominates when all agents carry their own unique reputation as in Whitaker et al. (2016). Prioritizing interaction with those who share the same trait (i.e. high  $s$ ) accelerates



**Figure 1.** Alternative agent-trait relationships for single-trait and multi-trait agents. (a)  $|G_1|=4$  single-trait dependent agents who share trait 1. (b) One multi-trait dependent agent, and five single-trait dependent agents.



**Figure 2.** Figure (a) shows the relationship between cooperation, parameter  $s$ , and the size of the set  $G_1$  of agents sharing a common trait (see Figure 1(a)). Figure (b) shows the effect of increasing the size of the set of traits  $T_1$  of a single multi-trait agent on cooperation, in a scenario where all other agents are single trait (see Figure 1(b)). (a) Dependent single-trait agents. (b) Dependent multi-trait agent.



**Figure 3.** Distribution of action rules (0, 0, 0) and (1, 1, 0) by generation for the sets of single-trait dependent agents  $G_1$  and independent agents  $A - G_1$ .  $|G_1| = 10$  and  $s=0$  (left), 0.5 (middle), and 1 (right).

the collapse of cooperation further as the discriminative strategy directs donations towards agents with a similar reputation. When  $s$  is low, dependent single-trait agents interact mainly with those who don't share their reputation as they are still incentivised to adopt cooperative strategies to maximize their fitness with a reduced risk of exploitation.

#### 4.2. Dependent multi-trait identities

In this section, we consider the effects of introducing a single dependent multi-trait agent (agent 1) in a population of single-trait agents. The schema for this arrangement is shown in Figure 1(b) and we vary  $|T_1|$ . The results (Figure 2(b)) show that as the number of traits held by agent 1 increases (i.e.  $|T_1|$ ), cooperation diminishes. This occurs similarly whether or not agent 1 plays with those who have at least one trait in common, as governed by  $s$ . The sharing agent 1's reputation is dispersed across single-trait agents that between

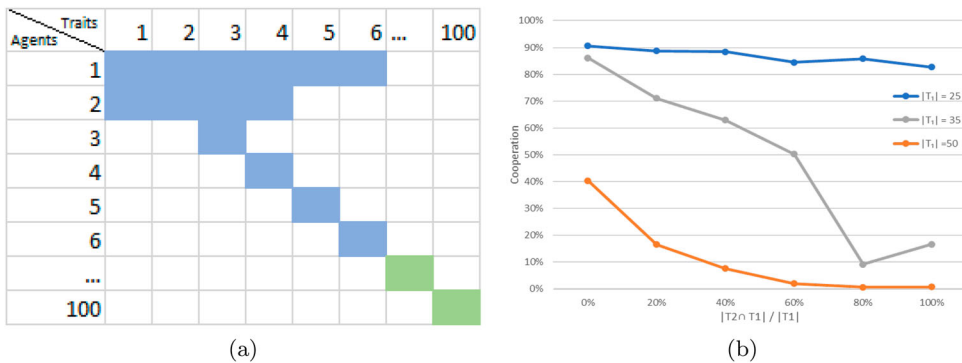
themselves have no trait in common. This helps to suppress the rise of defective action rules, as compared to the previous scenario (Section 4.1). In fact,  $|T_1|$  can reach a considerable size (e.g. 30–35 traits) before which cooperation starts to significantly diminish.

In this scenario, single-trait dependent agents rely entirely on themselves and the multi-trait agent for their reputation. Each single-trait dependent agent can also free ride on the single multi-trait agent, and this opens the opportunity for defection to establish itself, although to a lesser extent than the case presented in Section 4.1. When the number of traits of the multi-trait agent is relatively small, the presence of free-riding dependent single-trait agents can be sustained without too much disruption to the reputation of the multi-trait agent. As  $|T_1|$  increases, and the number of dependent single-trait agents increases, there is a greater opportunity for free-riding action rules to take hold (e.g.  $H_i = (0, 0, 0)$ ). At the same time, there are fewer independent single-trait agents available in the population. This promotes the collapse of cooperation. As soon as a defective strategy takes hold across the population, it then opens the opportunity for this to spread to other agents. Interestingly,  $s$  has relatively little impact on whether dependent agents prioritize playing with those that have a common trait. However, they are less likely to have an equal reputation in this instance.

Finally, we experiment with adding a second multi-trait agent, by replacing a single-trait agent (agent number 2) in Figure 1(b), where  $T_2 \subseteq T_1$ . Figure 4 shows the effect of varying  $|T_2 \cap T_1|$ , that is the extent to which  $T_2$  has the same traits as  $T_1$ . These results show that high proportions of shared identity through multi-trait agents undermine the reputation system. Because the second multi-trait agent can hold a large subset of the first agent's traits, it can heavily disrupt the first agent's reputation, by using defection as its action rule. This effect is more pronounced than that of a dependent single-trait agent sharing reputation with the multi-trait dependent agent, and increases as  $|T_2 \cap T_1|$  increases.

### 4.3. The evolution of identity

The previous experiments considered the evolution of behavioural action rules (strategies) while traits remained fixed throughout. In this section, we consider the effects of also

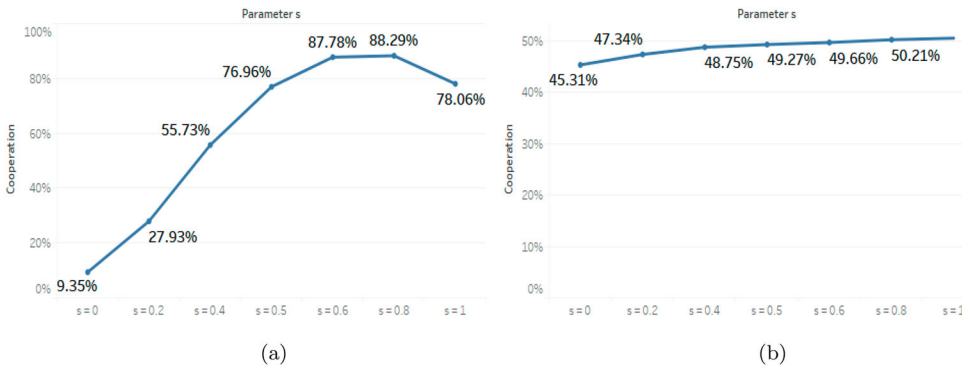


**Figure 4.** The figures show the relationship between agents and traits for two dependent multi-trait agents (left) and the average cooperation produced as a function of the size of the intersection between the sets belonging to agents one and two for different values of  $|T_1|$  where  $s=0$  (right).

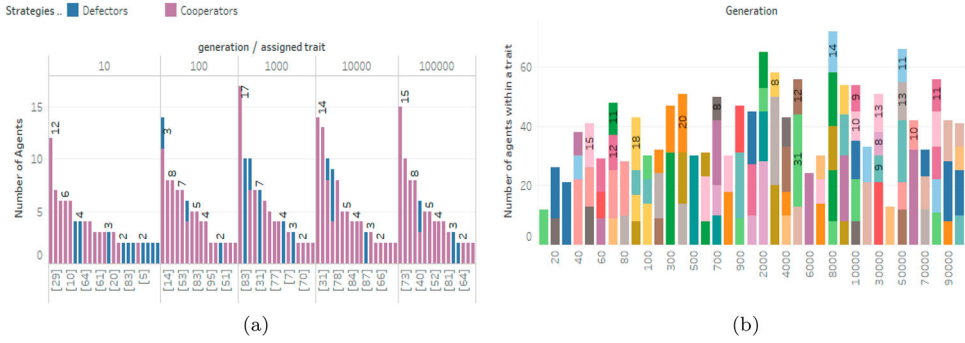
allowing agents to update their identity by changing their trait during the reproduction phase. To explore this scenario, we assume that each agent  $i$  has only a single trait ( $|T_i| = 1$ ), however, agents are able to share (and copy) the trait of another. The probability of an agent  $i$  changing to the identity of another agent  $j$  is proportional to  $j$ 's payoff relative to the whole population at the end of a generation. We further apply a mutation rate to the trait held by  $i$  to change into any other trait  $t \in T - \{t_i\}$ . Mutation of traits allows for strategies to arise in the population even when they have been removed through evolution. A trait mutation rate of 10% is applied, unless otherwise stated.

We experimented with a scenario where each individual initially has a single unique trait (referred to as 'independent agents'). However, because individuals are allowed to move between traits to promote payoff at the end of each generation, scenarios with different starting configurations have similar outcomes. The results in Figure 5(a) indicate that for the lowest values of  $s$  only limited cooperation is achieved, while it increases with higher rates of in-group interactions. Cooperation achieves an average of above 70% when individuals only interact with those having the same trait ( $s=1$ ). When  $s=0$  cooperation never reaches a level above 10% on average over 100,000 generations. This is in contrast with the outcomes obtained where identity remained fixed throughout the simulation, for which increasing the proportion of in-group interactions produced a sharp decrease in cooperative behaviour, see Figure 1.

When  $s=1$ , interactions of dependent agents are limited to agents who share their trait. Figure 6(b) shows that a trait can be shared between 10 and 30 agents before cooperation collapses, which is in line with previous experimentation (Figure 2(a)) where cooperation cannot be sustained when several agents share the same trait. The struggle for domination between the cooperators and defectors is seen in Figure 7(b). Here cooperative agents establish themselves with common identities and are then disrupted by defectors who adopt the same identity before they mutate to a new trait. This cycle repeats throughout the simulation (Figure 8(b)) and results in an average of above 70% cooperation. As evolution progresses shared traits increasingly tend to identify with cooperators (see Figure 6(a)).



**Figure 5.** The figures show the relationship between parameter  $s$  and cooperation when mutation on traits is applied at a rate of 10%. (a) Shows the relationship between cooperation and parameter  $s$  when the evolution of identity is enabled alongside the evolution of action rules. (b) Shows the relationship between parameter  $s$  and cooperation when agents are only allowed to evolve their identities but not their action rules.

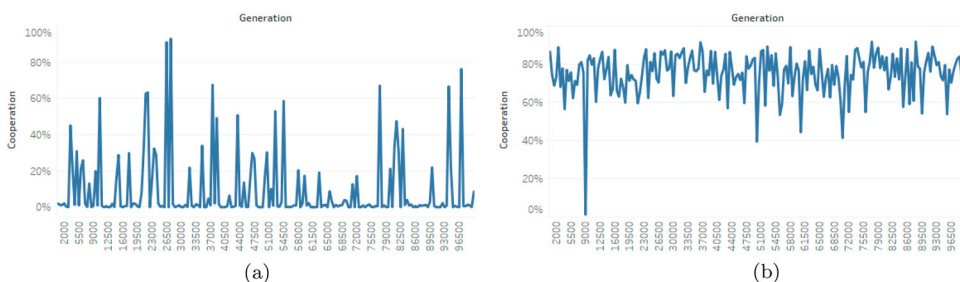


**Figure 6.** The frequency of agents that share a trait over generations when evolving both action rules and identity and  $s=1$ . (a) The frequency of strategies used by agents who are sharing traits at fixed generations (10 to 100,000). The figure shows that there are mostly cooperators than defectors within the most shared traits. (b) The most shared traits within each generation (where each different trait is represented by colour). The figure shows that a trait can be shared between 10 and 30 agents before cooperation collapses.

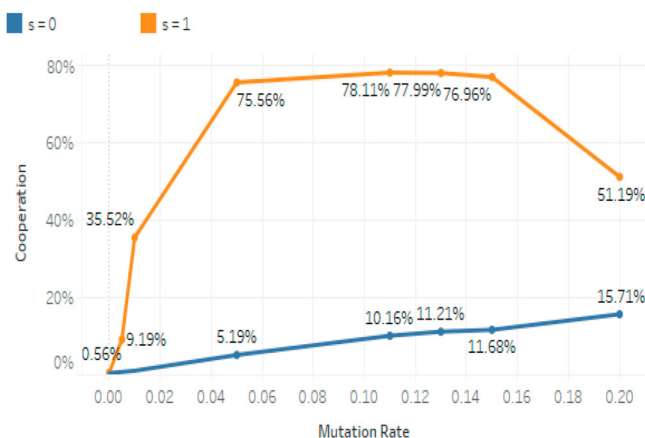


**Figure 7.** Distribution of strategies over generations when both strategies and identities are inherited. (a) When  $s=0$  the most frequent strategy within the population is the defector (0, 0, 0). This explains the lack of cooperation within the population that is displayed in Figure 8(a). (b) In contrast, when  $s=1$ , there is no clear dominant strategy within the population as it alternates between cooperators ( $s_i = 1$ ) and defectors ( $s_i = 0$ ) throughout the different generations.

In this scenario ( $s=1$ ), the reputation system becomes redundant because dependent agents only interact with those that have the same trait, and therefore the same reputation. This simplifies an agent's behaviour which becomes entirely dependent on  $s_i$  ( $u_i$  and  $d_i$  no longer function because all agents have the same reputation): simply  $i$  cooperates if  $s_i = 1$ , otherwise it defects. The set of interacting cooperators sharing the same trait maximize payoff and, as a result, attract other agents, increasing their number as long as their group doesn't involve defectors (i.e. agents  $i$  with  $s_i = 0$ ) that can benefit from shared reputation without donating. This provides opportunities for defective strategies to take hold and cooperation collapses. In this context, trait mutation allows cooperators to escape from defectors and move to an alternative trait. Figure 9 shows the criticality of mutation. When mutation is zero, agents are unable to escape from defectors. When the mutation rate is modestly increased (e.g. 1%), cooperative agents are able to change traits and rebuild a network of cooperative peers. However, when the mutation rate increases significantly, mutation impedes cooperation because



**Figure 8.** The pattern of cooperation over 100, 000 generations for  $s=0$  and  $s=1$  when evolving both identity and action rules. (a) Cooperation has a fluctuating trend throughout the 100, 000 generations when  $s=0$  producing a cooperation with an average below 10%. (b) Cooperation has a fluctuating trend throughout the 100, 000 generations when  $s=1$  producing a cooperation with an average above 70%.

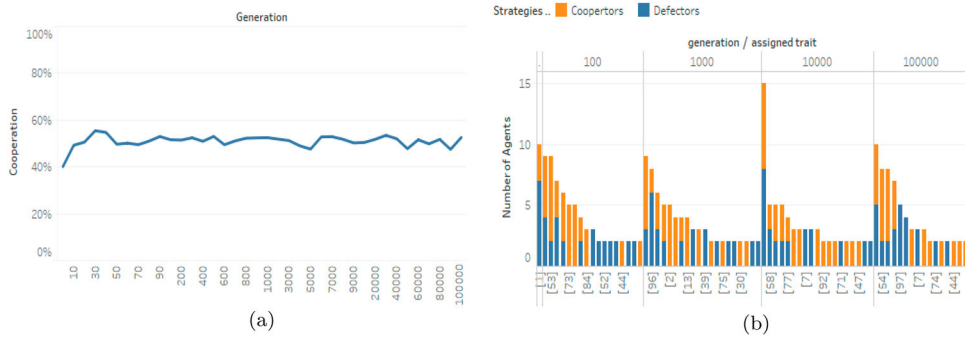


**Figure 9.** Cooperation vs. mutation rate for different values of  $s$ . The application of mutation on traits affects cooperation when  $s=1$  depending on the mutation rate applied. However, when  $s = 0$  mutation has a lower effect.

agents are rapidly mixing, increasing the chances of defectors and cooperators to share a trait (e.g. mutation of 100% is pure chance). This mechanism underlies the results in Figure 6, which shows how traits are shared by agents, with a few traits achieving a large amount of sharing by cooperators. These results align with the conclusions of Fu et al. (2012) and Tarnita et al. (2009), where a limited rate of trait mutation allows cooperators to rebuild, albeit improving on the levels of cooperation achieved. Similar techniques aimed to desert defectors as a mean to promote cooperation in absence of reputation, punishment, or other ostracizing mechanisms have also had relevance in the literature, see Aktipis (2004).

We note that this phenomenon only occurs when  $s=1$ , and allows for cooperation to be sustained in the presence of defectors. A similar behaviour is observed for high proportions of in-group interaction. In contrast, with high proportions of interaction outside the shared group ( $s=0$ ) and the reputation system fully in place, the same stereotyping effect described in Section 4.1 occurs, and cooperators do not establish themselves





**Figure 10.** Evolving identity rather than action rules. (a) When evolving identity without action rules cooperation presents small fluctuations throughout the 100, 000 generations producing an average of 50%, when  $s=1$ . (b) The most frequent strategies used within shared traits at specific generations (10–100, 000) with  $s = 1$ .

without exploitation as the groups sharing identity become increasingly large. This produces fluctuating cycles around lower cooperative levels (Figure 8(a)) leading to a substantial reduction in the average cooperation even in the presence of trait mutation, see Figure 9.

#### 4.3.1. Evolving identity rather than action rules

In this section we allow agents to evolve their identity while assuming that their action rules remain fixed throughout. In all other respects, we retain the parameter settings used in Section 4.3.

In these circumstances, the parameter  $s$  has a low impact on cooperation, as shown in Figure 5(b), where modest levels of cooperation are sustained (around 45–50%). As agents move towards interacting only with other agents who share the same trait, cooperation increases slightly. Changing identity does not offer protection against those holding defective strategies, as success equally attracts both defectors and cooperators to change identity. Accordingly, when a trait is associated with a healthy payoff, there is a likelihood of this being undermined by defectors in subsequent generations (Figure 10(a)) albeit with less fluctuation than when action rules are not fixed (Figure 8(b)).

## 5. Discussion and conclusion

In this work, we have addressed the issue of identity in cooperation systems, specifically concerning indirect reciprocity. Through a trait-based model, where traits hold reputation in their own right and can be shared between individuals, we allow reputation to be shared and combined. This challenges the current default assumption in modelling cooperation systems, which typically involves a one-to-one mapping between agents and their reputation. The results have established that cooperation can be heavily disrupted by the sharing of identity, when conditions allow for defective strategies to propagate through identity sharing. We note that our model does not involve any secondary mechanisms aligned to group-based social norms (i.e. human kinship), which may function

to promote cooperation in a particular group without reputation and where there is a common identity, as in generalized reciprocity.

In our general scenario, stereotyping takes place where common traits are used as proxy an individuals identity and their reputation. This introduces the opportunity for agents to disconnect their actions from their reputation. Through this mechanism, agents can deploy defective strategies: that is an agent can avoid paying the full costs of donations, but still receive them based on the reputation aligned with its associated traits. How identity is shared, through inheritance of traits, is highly influential. Holding multiple traits presents an opportunity for agents to share a limited proportion of their identity with others. In doing so they have the potential to better control their exposure to defectors.

Single-trait and multi-trait agents are differentiated in how they share traits with others. Under uniform conditions, single-trait agents have a reduced chance of having a trait in common. However, when another agent shares their trait, their reputation becomes susceptible to the actions of a third party. In contrast, for multi-trait agents, increasing the number of traits can give them a chance to retain an element of unique personal identity, through traits that aren't shared with others. Moreover, for multi-trait agents, sharing can occur with a number of agents that have no dependency between them, in terms of common traits. The results show that reasonable levels of cooperation can be sustained while there is a modest level of sharing of identity in the population, after which cooperation collapses.

We have also examined the consequences of identity change becoming an element of an agent's strategy. This extends the concept of whitewashing, allowing agents to legitimately share traits with those that are successful. The results show that this can be damaging to the emergence of cooperation. However, when conditions are imposed that dictate agents should primarily play with those having the same trait in common, significant cooperation emerges. We have found that this creates an interesting set of conditions where the reputation system collapses, the agents are divided into mutually exclusive sets that are identified by traits, and donation decisions are made on a single component of an agent's action rule ( $s_i$ ). In these circumstances, the presence of a modest amount of trait mutation is sufficient to allow cooperation to emerge. The results indicate bursty-ness in cooperation as evolution takes place, due to payoff through cooperators with a common trait being subsequently undermined by the presence of agents with defective strategies. This finding is significant because it shows how a reputation-based cooperation systems collapses into set-based evolution (Fu et al., 2012; Tarnita et al., 2009), which bridges alternative perspectives on evolution other than indirect reciprocity (i.e. generalized reciprocity, see Yamagishi & Kiyonari, 2000).

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

W. Bedewi is funded by King Abdulaziz University, Saudi Arabia. Additionally, the research was sponsored by the U.S. Army Research Laboratory and the U.K. Ministry of Defence under Agreement

Number W911NF-16-3-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory, the U.S. Government, the U.K. Ministry of Defence or the U.K. Government. The U.S. and U.K. Governments are authorized to reproduce and distribute reprints for Government purposes not withstanding any copyright notation hereon. This research was also supported by the Supercomputing Wales project, which is part-funded by the European Regional Development Fund (ERDF) via Welsh Government.

## Notes on contributors

**Wafi Bedewi** is a PhD student at Cardiff University, UK, studying the role of identity in cooperation systems through simulation-based methodologies. He also holds a Lecturer position at King Abdulaziz University, Jeddah, Saudi Arabia. Wafi received his MSc degree in Information Security and Privacy in 2015 from Cardiff University.

**Roger M. Whitaker** is a Professor of Collective Intelligence with Cardiff University, UK, with interests in evolutionary processes, cooperation and intelligence. He is currently the Dean of Research and Innovation of the College of Physical Sciences and Engineering and the Director of Supercomputing Wales, the national facility for high performance computing in Wales.

**Gualtiero B. (Walter) Colombo** is a Research Software Engineer in support of Supercomputing Wales, UK. Walter has a background in structural civil engineering and computing (meta-heuristics) and is experienced in interdisciplinary research, particularly related to genetic algorithms, networks and social processes including cooperation and multi-agent systems.

**Stuart M. Allen** is currently a Professor and the Head of the School of Computer Science and Informatics at Cardiff University, UK. From a background in discrete mathematics and optimization, his research interests are in the area of mobile and social computing. He currently serves on the editorial boards for Computer Communications and the Proceedings of the Royal Society A.

**Yarrow Dunham** is an Assistant Professor of psychology and cognitive science and the Director of the Social Cognitive Development Lab at socialcogdev.com. His research focuses on studying how knowledge of social groups is acquired, both in cognitively mature adults and in developing children. His research combines a range of experimental and cross-cultural methodologies.

## ORCID

Wafi Bedewi  <http://orcid.org/0000-0001-6450-7968>

Roger M. Whitaker  <http://orcid.org/0000-0002-8473-1913>

Stuart M. Allen  <http://orcid.org/0000-0003-1776-7489>

## References

- Aktipis, C. A. (2004). Know when to walk away: Contingent movement and the evolution of cooperation. *Journal of Theoretical Biology*, 231(2), 249–260. <https://doi.org/10.1016/j.jtbi.2004.06.020>
- Alexander, R. D. (1987). *The biology of moral systems*. Transaction.
- Baranski, B., Bartz-Beielstein, T., Ehlers, R., Kajendran, T., Kossler, B., Mehnen, J., Polazek, T., Reimholz, R., Schmidt, J. M., Schmitt, K., Seis, D., Slodzinski, R., Steeg, S., Wiemann, N., & Zimmermann, M. (2006). *The impact of group reputation in multiagent environments*. IEEE International Conference on Evolutionary Computation, Vancouver, BC, Canada (pp. 1224–1231). <https://doi.org/10.1109/CEC.2006.1688449>.
- Bear, A., & Rand, D. G. (2016). Intuition, deliberation, and the evolution of cooperation. *Proceedings of the National Academy of Sciences*, 113(4), 936–941. <https://doi.org/10.1073/pnas.1517780113>

- Bowleg, L. (2017). Intersectionality: An underutilized but essential theoretical framework for social psychology. In B. Gough (Ed.), *The Palgrave handbook of critical social psychology* (pp. 507–529). Palgrave Macmillan. [https://doi.org/10.1057/978-1-137-51018-1\\_25](https://doi.org/10.1057/978-1-137-51018-1_25)
- Brandt, H., Ohtsuki, H., Iwasa, Y., & Sigmund, K. (2007). A survey of indirect reciprocity. In Y. Takeuchi, Y. Iwasa, & K. Sato (Eds.), *Mathematics for ecology and environmental sciences. Biological and medical physics, biomedical engineering* (pp. 21–49). Springer. [https://doi.org/10.1007/978-3-540-34428-5\\_3](https://doi.org/10.1007/978-3-540-34428-5_3)
- Brandt, H., & Sigmund, K. (2004). The logic of reprobation: Assessment and action rules for indirect reciprocation. *Journal of Theoretical Biology*, 231(4), 475–486. <https://doi.org/10.1016/j.jtbi.2004.06.032>
- Brewer, M. B. (1999). The psychology of prejudice: Ingroup love and outgroup hate? *Journal of Social Issues*, 55(3), 429–444. <https://doi.org/10.1111/josi.1999.55.issue-3>
- de Melo, C. M., Marsella, S., & Gratch, J. (2019). Human cooperation When acting through autonomous machines. *Proceedings of the National Academy of Sciences*, 116(9), 3482–3487. <https://doi.org/10.1073/pnas.1817656116>
- Dovidio, J. F., Gaertner, S. L., & Validzic, A. (1998). Intergroup bias: Status, differentiation, and a common in-group identity. *Journal of Personality and Social Psychology*, 75(1), 109. <https://doi.org/10.1037/0022-3514.75.1.109>
- Fehr, E. (2004). Human behaviour: Don't lose your reputation. *Nature*, 432(7016), 449–450. <https://doi.org/10.1038/432449a>
- Feldman, M., & Chuang, J. (2005a). *The evolution of cooperation under cheap pseudonyms*. Seventh IEEE International Conference on E-Commerce Technology (CEC'05), Munich, Germany (pp. 284–291). IEEE. <https://doi.org/10.1109/icect.2005.91>
- Feldman, M., & Chuang, J. (2005b). Overcoming free-riding behavior in peer-to-peer systems. *ACM Sigecom Exchanges*, 5(4), 41–50. <https://doi.org/10.1145/1120717>
- Fishman, M. A. (2003). Indirect reciprocity among imperfect individuals. *Journal of Theoretical Biology*, 225(3), 285–292. [https://doi.org/10.1016/S0022-5193\(03\)00246-7](https://doi.org/10.1016/S0022-5193(03)00246-7)
- Fu, F., Tarnita, C. E., Christakis, N. A., Wang, L., Rand, D. G., & Nowak, M. A. (2012). Evolution of in-group favoritism. *Scientific Reports*, 2, 460. <https://doi.org/10.1038/srep00460>
- Galinsky, A. D., & Moskowitz, G. B. (2000). Perspective-taking: Decreasing stereotype expression, stereotype accessibility, and in-group favoritism. *Journal of Personality and Social Psychology*, 78(4), 708. <https://doi.org/10.1037/0022-3514.78.4.708>
- Gao, S., Wu, T., & Wang, L. (2018). Evolution of global cooperation and ethnocentrism in group-structured populations. *Physics Letters A*, 382(31), 2027–2043. <https://doi.org/10.1016/j.physleta.2018.05.020>
- Hales, D. (1998). Stereotyping, groups and cultural evolution: A case of “second order emergence”? In J. S. Sichman, R. Conte, & N. Gilbert (Eds.), *Multi-agent systems and agent-based simulation* (pp. 140–155). Springer-Verlag. [https://doi.org/10.1007/10692956\\_10](https://doi.org/10.1007/10692956_10)
- Hamilton, W. D. (1964). The genetical evolution of social behaviour. II. *Journal of Theoretical Biology*, 7(1), 17–52. [https://doi.org/10.1016/0022-5193\(64\)90039-6](https://doi.org/10.1016/0022-5193(64)90039-6)
- Hammond, R. A., & Axelrod, R. (2006). The evolution of ethnocentrism. *Journal of Conflict Resolution*, 50(6), 926–936. <https://doi.org/10.1177/0022002706293470>
- Hoffman, K., Zage, D., & Nita-Rotaru, C. (2009). A survey of attack and defense techniques for reputation systems. *ACM Computing Surveys (CSUR)*, 42(1), 1. <https://doi.org/10.1145/1592451>
- Imbsweiler, J., Ruesch, M., Weinreuter, H., León, F. P., & Deml, B. (2018). Cooperation behaviour of road users in t-intersections during deadlock situations. *Transportation Research Part F: Traffic Psychology and Behaviour*, 58, 665–677. <https://doi.org/10.1016/j.trf.2018.07.006>
- Kawakami, K., Amodio, D. M., & Hugenberg, K. (2017). Intergroup perception and cognition: An integrative framework for understanding the causes and consequences of social categorization. In J. M. Olson (Ed.), *Advances in experimental social psychology* (Vol. 55, pp. 1–80). Elsevier Academic Press.
- Launay, J., & Dunbar, R. I. M. (2015). Playing with strangers: Which shared traits attract us most to new people? *PLoS One*, 10(6), 1–17. <https://doi.org/10.1371/journal.pone.0129688>

- Leimar, O., & Hammerstein, P. (2001). Evolution of cooperation through indirect reciprocity. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 268(1468), 745–753. <https://doi.org/10.1098/rspb.2000.1573>
- Li, K., Cong, R., & Wang, L. (2016). Stochastic evolutionary dynamics in minimum-effort coordination games. *Physics Letters A*, 380(34), 2595–2602. <https://doi.org/10.1016/j.physleta.2016.06.007>
- Lieberman, E., Hauert, C., & Nowak, M. A. (2005). Evolutionary dynamics on graphs. *Nature*, 433(7023), 312. <https://doi.org/10.1038/nature03204>
- Masuda, N. (2012). Ingroup favoritism and intergroup cooperation under indirect reciprocity based on group reputation. *Journal of Theoretical Biology*, 311, 8–18. <https://doi.org/10.1016/j.jtbi.2012.07.002>
- Melnik, M. I., & Alm, J. (2002). Does a seller's ecommerce reputation matter? Evidence from eBay auctions. *The Journal of Industrial Economics*, 50(3), 337–349. <https://doi.org/10.1111/1467-6451.00180>
- Milinski, M., Semmann, D., & Krambeck, H. J. (2002). Reputation helps solve the 'tragedy of the commons'. *Nature*, 415(6870), 424–426. <https://doi.org/10.1038/415424a>
- Molleman, L., van den Broek, E., & Egas, M. (2013). Personal experience and reputation interact in human decisions to help reciprocally. *Proceedings of the Royal Society B: Biological Sciences*, 280(1757), 20123044. <https://doi.org/10.1098/rspb.2012.3044>
- Nathanson, C. G., Tarnita, C. E., & Nowak, M. A. (2009). Calculating evolutionary dynamics in structured populations. *PLoS computational biology*, 5(12), e1000615. <https://doi.org/10.1371/journal.pcbi.1000615>
- Nowak, M. A. (2006). Five rules for the evolution of cooperation. *Science*, 314(5805), 1560–1563. <https://doi.org/10.1126/science.1133755>
- Nowak, M. A., & May, R. M. (1992). Evolutionary games and spatial chaos. *Nature*, 359(6398), 826. <https://doi.org/10.1038/359826a0>
- Nowak, M. A., & Sigmund, K. (1998). Evolution of indirect reciprocity by image scoring. *Nature*, 393(6685), 573–577. <https://doi.org/10.1038/31225>
- Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature*, 437(7063), 1291–1298. <https://doi.org/10.1038/nature04131>
- Nowak, M. A., Tarnita, C. E., & Wilson, E. O. (2010). The evolution of eusociality. *Nature*, 466(7310), 1057. <https://doi.org/10.1038/nature09205>
- Oakes, P. J., & Turner, J. C. (1980). Social categorization and intergroup behaviour: Does minimal intergroup discrimination make social identity more positive? *European Journal of Social Psychology*, 10(3), 295–301. [https://doi.org/10.1002/\(ISSN\)1099-0992](https://doi.org/10.1002/(ISSN)1099-0992)
- Ohtsuki, H., & Iwasa, Y. (2006). The leading eight: Social norms that can maintain cooperation by indirect reciprocity. *Journal of Theoretical Biology*, 239(4), 435–444. <https://doi.org/10.1016/j.jtbi.2005.08.008>
- Panchanathan, K., & Boyd, R. (2003). A tale of two defectors: The importance of standing for evolution of indirect reciprocity. *Journal of Theoretical Biology*, 224(1), 115–126. [https://doi.org/10.1016/S0022-5193\(03\)00154-1](https://doi.org/10.1016/S0022-5193(03)00154-1)
- Pinker, S. (2012). The false allure of group selection. <http://edge.org/conversation/the-false-allure-of-group-selection>
- Rand, D. G., & Nowak, M. A. (2013). Human cooperation. *Trends in Cognitive Sciences*, 17(8), 413–425. <https://doi.org/10.1016/j.tics.2013.06.003>
- Resnick, P., Kuwabara, K., Zeckhauser, R., & Friedman, E. (2000). Reputation systems. *Communications of the ACM*, 43(12), 45–48. <https://doi.org/10.1145/355112.355122>
- Riolo, R. L., Cohen, M. D., & Axelrod, R. (2001). Evolution of cooperation without reciprocity. *Nature*, 414(6862), 441. <https://doi.org/10.1038/35106555>
- Smith, J. M. (1964). Group selection and kin selection. *Nature*, 201(4924), 1145. <https://doi.org/10.1038/201145a0>
- Sugden, R. (1986). *The economics of rights, co-operation and welfare*. Blackwell.
- Suzuki, S., & Akiyama, E. (2005). Reputation and the evolution of cooperation in sizable groups. *Proceedings of the Royal Society B: Biological Sciences*, 272(1570), 1373–1377. <https://doi.org/10.1098/rspb.2005.3072>

- Swann, Jr., W. B., & M. D. Buhrmester (2015). Identity fusion. *Current Directions in Psychological Science*, 24(1), 52–57. <https://doi.org/10.1177/0963721414551363>
- Tajfel, H. (1974). Social identity and intergroup behaviour. *Information (International Social Science Council)*, 13(2), 65–93. <https://doi.org/10.1177/053901847401300204>
- Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971). Social categorization and intergroup behaviour. *European Journal of Social Psychology*, 1(2), 149–178. [https://doi.org/10.1002/\(ISSN\)1099-0992](https://doi.org/10.1002/(ISSN)1099-0992)
- Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. In W. G. Austin & S. Worchel (Eds.), *The Social psychology of inter-group relations* (pp. 33–47). Brooks/Cole.
- Takahashi, N., & Mashima, R. (2006). The importance of subjectivity in perceptual errors on the emergence of indirect reciprocity. *Journal of Theoretical Biology*, 243(3), 418–436. <https://doi.org/10.1016/j.jtbi.2006.05.014>
- Tarnita, C. E., Antal, T., Ohtsuki, H., & Nowak, M. A. (2009). Evolutionary dynamics in set structured populations. *Proceedings of the National Academy of Sciences*, 106(21), 8601–8604. <https://doi.org/10.1073/pnas.0903019106>
- Tarnita, C. E., Wage, N., & Nowak, M. A. (2011). Multiple strategies in structured populations. *Proceedings of the National Academy of Sciences*, 108(6), 2334–2337. <https://doi.org/10.1073/pnas.1016008108>
- Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., & Wetherell, M. S. (1987). *Rediscovering the social group: A self-categorization theory*. Basil Blackwell.
- Wasko, M. M., & Faraj, S. (2000). “It is what one does”: Why people participate and help others in electronic communities of practice. *The Journal of Strategic Information Systems*, 9(2-3), 155–173. [https://doi.org/10.1016/S0963-8687\(00\)00045-7](https://doi.org/10.1016/S0963-8687(00)00045-7)
- Wedekind, C., & Milinski, M. (2000). Cooperation through image scoring in humans. *Science*, 288(5467), 850–852. <https://doi.org/10.1126/science.288.5467.850>
- Whitaker, R. M., Colombo, G. B., Allen, S. M., & Dunbar, R. I. (2016). A dominant social comparison heuristic unites alternative mechanisms for the evolution of indirect reciprocity. *Scientific reports*, 6, 31459. <https://doi.org/10.1038/srep31459>
- Whitaker, R. M., Colombo, G. B., & Rand, D. G. (2018). Indirect reciprocity and the evolution of prejudicial groups. *Scientific Reports*, 8(1), 13247. <https://doi.org/10.1038/s41598-018-31363-z>
- Wilson, D. S. (1975). A theory of group selection. *Proceedings of the National Academy of Sciences*, 72(1), 143–146. <https://doi.org/10.1073/pnas.72.1.143>
- Wilson, D. S., & Sober, E. (1994). Reintroducing group selection to the human behavioral sciences. *Behavioral and brain sciences*, 17(4), 585–608. <https://doi.org/10.1017/S0140525X00036104>
- Wu, J., Balliet, D., & Van Lange, P. A. (2016). Reputation, gossip, and human cooperation. *Social and Personality Psychology Compass*, 10(6), 350–364. <https://doi.org/10.1111/spc3.v10.6>
- Yamagishi, T., & Kiyonari, T. (2000). The group as the container of generalized reciprocity. *Social Psychology Quarterly*, 63, 116–132. <https://doi.org/10.2307/2695887>