The Journal of Mind and Behavior
Summer and Autumn 2016, Volume 37, Numbers 3 and 4
Pages 247–270
ISSN 0271–0137

Reciprocity and Reputation: A Review of Direct and Indirect Social Information Gathering

Yvan I. Russell

University of Göttingen and *Middlesex University*

Direct reciprocity, indirect reciprocity, and reputation are important interrelated topics in the evolution of sociality. This non-mathematical review is a summary of each. Direct reciprocity (the positive kind) has a straightforward structure (e.g., "A rewards B, then B rewards A") but the allocation might differ from the process that enabled it (e.g., whether it is true reciprocity or some form of mutualism). Indirect reciprocity (the positive kind) occurs when person (B) is rewarded by a third party (A) after doing a good deed towards somebody else (C) — with the structure "A observes B help C, therefore A helps B." Here too, the allocation differs from the process: if there is underlying cognition, then indirect reciprocity is based on some ability to keep track of the reputations of others (to remember that "B helped C"). Reputation is a kind of social impression based on typicality, derived from three channels of experience (direct encounters, bystander observation, and gossip). Although non-human animals cannot gossip verbally, they can eavesdrop on third parties and learn vicariously. This paper ends with a proposal to investigate the topic of social expertise as a model for understanding how animals understand and utilise observed information within their social groups.

Keywords: reputation, reciprocity, cooperation, expertise

Reputation as an Animal Concept

In our daily lives, we often cogitate on matters of gossip and reputation. People get upset over a bad reputation. This was humorously illustrated in Anton Chekhov's 1883 short story "A Slander" ("Клевета"), whereupon a prestigious schoolmaster, attending his daughter's wedding, smacks his lips in approval of some delicious food prepared in the kitchen — and then later is obsessively chagrined after discovering that his innocent lip-smacking noise was heard by someone in an adjoining room, leading to a widespread rumour that he was adulterously kissing the female

I would like to thank two anonymous reviewers for their helpful comments. I would also like to thank Professor Raymond Russ (University of Maine) for extensive feedback and guidance. Correspondence concerning this article should be addressed to Dr. Yvan I. Russell, Department of Psychology, Middlesex University, The Burroughs, London NW4 4BT, United Kingdom, Email: yvanrussell@gmail.com

cook who was in the same room when he made the noise (Chekhov, 1883/1921). Stories like this illustrate "gossip" as a pejorative term, something that might unjustly cause embarrassment and suffering.

However, we can use the word "gossip" in a non-pejorative sense too: simply as a mechanism of social information exchange. Most human conversations are dominated by social gossip, which suggests that gossip has important functions (Dunbar, 2004). Gossip "allows individuals and communities to accumulate behavioural evidence about others and to form and refine judgements about their vices and virtues" (Emler, 1994, p. 133). In other words, gossip gives us important information. If a person has a bad reputation, then one might be inclined to avoid that person (lest one suffer the way that others have suffered). If a person has a good reputation, then one might be inclined to approach that person (to benefit the way that others have benefited). If you know nothing about a person's reputation, then you approach that person as a blank slate with no predictive information on whether you will encounter positive, neutral, or negative consequences. If the stakes are high, then it is useful to take advantage of information gathered by others. Here, we can bring in a biological concept: the producer–scrounger effect, where the thief (scrounger) takes advantage of the "behavioural investment of another (producer) to obtain a limited resource" (Barnard and Sibly, 1981, p. 543). Usually, this concept is applied to phenomena such as kleptoparisitism (stealing food from one who made the — perhaps risky — effort to acquire the food; this is a low-effort way to obtain food, and a loss for the other, e.g., Spencer, Russell, Dickins, and Dickins, 2017). We can apply the concept of producer–scrounger to social information gathering, where the "limited resource" is information. Imagine that you need information that will help decide whether to approach a man called Mr. Enemy. Imagine further that you witnessed Mr. Enemy injuring Mr. Friend. Here, Mr. Friend made the — perhaps risky — effort to "produce" information for you. By "scrounging" information produced by Mr. Friend (seeing him get injured), you have gained valuable information (that you should perhaps avoid Mr. Enemy) whilst avoiding personal injury yourself. There is an advantage to gathering information by proxy. The cognitive mechanism here is analogical reasoning: "if Mr. Enemy hurt Mr. Friend, then he will probably hurt me too." I return to the topic of "if–then" social reasoning (de Waal, 2003) later.

Like others before me (e.g., Dunbar, 2004), I am interested in gossip and reputation in an evolutionary context. I view it as important to adapt the concept of reputation in such a way as to accommodate the capabilities of the whole animal kingdom (Russell, 2007). The word "reputation" is normally used to describe an exclusively human activity: "person A gives testimony about person B to person C, teaching C something new about B." In everyday colloquial usage, "reputation" refers to this transitivity: information verbally passing along the gossip network from one person to another. Language is the crucial ingredient of this kind of information flow, and without it, information does not flow past those

who perceived the signals first-hand. Verbal information becomes highly advantageous for cooperation when the size of a social group exceeds the point where an individual can rely on first-hand knowledge (see Greif, 1989, for an historical example). A good reputation is beneficial. As Alexander (1987, p. 95) wrote: "The concept of status implies that an individual's privileges, or its access to resources, are controlled in part by how others collectively think of him (hence, treat him) as a result of past interactions (including observations of interactions with others)." With this advantage in mind, it makes sense that people are motivated to behave well when they know they are being watched (Emler, 1990; Engelmann and Fischbacher, 2009). To adapt the concept of reputation for non-human animals, I start by recognising that reputation involves more than language. Information about others can be also gained from sheer observation (well within the capabilities of animals). Therefore, I have previously (Russell, 2007; Russell, Call, and Dunbar, 2008) defined reputation in an animal-inclusive manner as *knowledge about an individual's typical behaviour based on a knowledge of that individual's past behaviour*.

In many situations, the word "individual" can be replaced with "entity," because people routinely assign reputations to groups of individuals (even whole nations), corporate entities in the business world, or even insentient objects and phenomena. The attribution of reputations to groups is in several ways a parallel (but not identical) process to the attribution of stereotypes (cf. McGarty, 2002 and Spears, 2011). Reputation has been defined in many diverse ways by a large number of different researchers (e.g., see the reputation model developed for the business world by Carmeli and Tishler, 2005, who define reputation as the external perception of a business in terms of distinctiveness and prestige). The word "reputation" itself (like many words) can seem semantically opaque about its actual referent. It is not really a concrete "thing," but rather it is information which somehow has a life of its own beyond individual brains. Here, I focus on *individual* reputations. Reputations are not an inherent property of an individual, but are subjective attributions made by others (Obreiter, Fähnrich, and Gianluca, 2005; Pollock and Dugatkin, 1992). Among non-human animals, reputation can only exist completely outside the sphere of language. Animals respond deftly to motivational cues and signals from other animals (Krebs and Dawkins, 1984; Smith and Harper, 2004), but reputation is relevant only if knowledge about a particular individual's past behaviour is remembered and influences current behaviour towards that individual. Animals may learn the typical behaviour of others in three ways (Smith and Harper, 2004): (1) direct reputation (personal encounters), (2) indirect reputation (observing events as uninvolved bystander), and (3) reported reputation (gossip) [cf; Ostrom, 2003, pp. 43-44]. Whilst verbal gossip is surely uniquely human, the other channels of information (direct reputation and indirect reputation) are usable by animals to varying extents (depending on the species and its cognitive abilities; see discussion in Russell, 2007).

The actual content of human gossip varies widely. In humans, it is often about personality traits (such as Chekhov's unfortunate schoolmaster being labelled an adulterer) or about episodes that lead to personality attributions (that same schoolmaster being heard to make a kissing sound). Social psychological research on the topic of reputation has tended to focus on information related to personality attributes (Emler, 1990). In order to apply the concept of reputation to animals, we need to focus on more basic behaviour. Below, I will focus on the most basic "moral" behaviours. In doing so, I adopt the view that biology and morality are intertwined (cf. Alexander, 1987). Theorists such as Alexander (1987) and Binmore (2005) promote an empirical and naturalistic view of morality: instead of prescribing rules based on abstract principles, we can study what humans actually do. We can view morality as being based on social contracts, and success measured by the establishment of *equilibria* (see below about the Nash equilibrium). However, as Hamilton (1975) wrote, contractual morality has an in-built uncertainty: "It is very frequently necessary for one party to execute his half of a bargain without any way of being certain that the other party will later stick to his" (p. 150). Below, I review the concepts of direct reciprocity, indirect reciprocity, and reputation as mechanisms that help ensure that the second half of the bargain is met.

Direct Reciprocity

There are many forms of dyadic (two-person) reciprocity (see Dugatkin, 1997), but here I describe the most basic form (Binmore, 2005; Dugatkin, 1997; Sigmund, 2010): "When individual A copies what individual B does. Hence, if B gives, A gives back; if B fails to give, then A defects in return" (failure to reciprocate is called a "defection"). If a reciprocal relationship lasts for multiple rounds, then it can take the appearance of a feedback circuit or loop: A pays B, then B pays A back. Kolm (2000) averred that reciprocity is classifiable in two different ways: according to allocation (the actual budget of given and received items) and process (the mechanism that enabled it — for example a psychological motivation). In this paper, we will reflect on both allocation and process.

Stable cooperative relationships are formed as a summation of repeated interactions, the exact pattern of which is unique to a particular dyad (Bowles and Gintis, 2011; Hinde, 1976). Mutually beneficial relationships are established after a progressive reduction of uncertainty between two actors about the benefits that arise after one actor signals an intention to benefit the other (Markl, 1985). A feedback circuit is perhaps not the best metaphor, because being a sender/receiver is a role (rather than a characteristic) of an organism (Markl, 1985). For both parties to continue to respond to each other, there needs to be some two-way payoff; otherwise, "nothing in the

world can keep receivers dancing like puppets on the strings of the senders' signals — unless it is to their own advantage, too, to be manipulated" (Markl, 1985, p. 165). Responsiveness can be called "tightness" (Markl, 1985): how tight the loop is between signal and response (and whether a response occurs at all). When direct (pairwise) reciprocity occurs between sentient animals in natural settings, the "circuit" is created through behavioural episodes occurring at fairly unpredictable intervals between organisms who may or may not transact again. Each dyad, furthermore, does not stand in isolation, but is embedded within the complexity of an ecological niche with its connate social network (Alexander, 1987; Clutton–Brock, 2009; Hinde, 1976; McGregor, 2005; Nowak and Highfield, 2011; Ostrom, 2003). Reciprocation does not necessarily consist of costs and benefits for each side; benefits may flow with both parties gaining rather than losing, becoming what is termed pseudoreciprocity or by-product mutualism (Alexander, 1987; Clutton–Brock, 2009; Dugatkin, 1997). It is useful to focus on the simpler constituents of prosociality as a step towards understanding the larger social/cooperative structure of animal and human societies (Alexander, 1977, 1987; Clutton–Brock, 2009; Dawes, 1980; Dugatkin, 1997; Hinde, 1976; McGregor, 2005; Nowak and Highfield, 2011; Sigmund, 2010, etc.). Reciprocity is widely regarded as a key mechanism in human sociality (Fehr and Gächter, 2000; Kolm, 2000). Episodes of reciprocation — so common across cultures, between friends or strangers, ingrained in our social norms — are elementary units of our cooperative societies which provide public goods, and which contribute to the survival of members of our species (Alexander, 1987; Dawes, 1980; Nowak and Highfield, 2011).

Markl (1985, p. 170) identified four scenarios of payoffs for dyadic relationships: (1) both actors benefit (cooperation), (2) the sender benefits but not the receiver (exploitative), (3) the receiver benefits but not the sender (also exploitative), and (4) neither benefit (a wasted effort). Direct reciprocity falls into the first category, but only if reciprocation occurs (otherwise it is exploitation). We can analyse these relationships using *game theory*: where a player's probability of payoff is contingent on the behaviour of others (see Binmore, 2005; Dugatkin, 1997; Sigmund, 2010). The colloquialism "I'll scratch your back, you scratch mine" is often invoked as a one-liner summary of reciprocity. To prevent oneself from suffering defection, it is helpful to avoid one-shot encounters and benefit from repeated encounters with reliable individuals. Binmore (2005, p. 10) elaborated:

> Rational reciprocity can't work unless people interact repeatedly, without a definite end to their relationship in sight. If the reason I scratch your back today is that I expect you will then scratch my back tomorrow, then our cooperative arrangement will unravel if we know that there will eventually be no tomorrow.

However, the possible extent of calculation is limited. Imagine that a human being (the scratcher) has his emotional life and culture all stripped away and what

remains is a coldly calculating person who thinks only of the payoff. For example, this robotic-type might say to himself: "if I scratch his back for ten minutes, then there is a 95% chance that he will scratch my back for ten minutes within a week." No person actually thinks quantitatively in this manner (precision would be impossible). However, we are influenced by this kind of rationalistic dynamic on a functional level. We might make the analogy of a rat in a Skinner box, being influenced by the principles of operant conditioning (e.g., Guttman, 1953). The actual quantitative explanation of the rat's behaviour is only calculable by the scientist standing outside the box taking measurements. The rat itself cannot understand the operant principles governing its own behaviour. We humans might regard ourselves intellectually superior to a rat — but psychologists (e.g., Simon, 1955, 1983) have known for a long time that cognitively we just do not follow the economic rules of "rational man"; instead, we put in just enough mental effort to attain some desired outcome (Gigerenzer, 1997) because we are not privy to the full information that would enable us to maximise our benefits at every step of our daily behaviour (Simon, 1983). Furthermore, we are highly imperfect reasoning machines, subject to numerous biases (Ayton, 2010). We humans are typically like the rat in the Skinner box, and this includes situations where we respond to costs and benefits of reciprocal interactions (Binmore, 2005; Ostrom, 2003), much like how the unreflective rat in the Skinner box responds to the benefits of pushing a lever. That non-human animals show reciprocation behaviours is well established, although the proximate mechanisms (processes) are debated (see Clutton–Brock, 2009).

Game theory is a system for investigating how payoffs differ according to the strategy adopted. Payoffs can be anything. For chimpanzees, payoff might literally be the receipt of "back-scratching" (i.e., social grooming, see Russell and Phelps, 2013). For humans who play economic games, the payoff might be money (Dawes, 1980). In direct reciprocity among moneyless organisms, the payoff might be your future reproductive success (Trivers, 1971). PAYOFF is a generic concept. Accordingly, game theorists refer to *utility* (and its unit of measurement, *util*), a generic unit of payoff that results from a given decision (Binmore, 2005; Simon, 1955): as a currency, an util can be anything (whether it's reproductive success, food, actual money, etc.) and even when undefined it can be used as a variable in biological or cognitive modelling (see Bowles and Gintis, 2011). In evolutionary game theory, costs and benefits can be numbered as "fitness units" addable or subtractable from a baseline fitness (Sober and Wilson, 1998). In our everyday thinking, we lack the perfect information that allows robotic-like total rationality — so instead we rely on our limited information and use cognitive shortcuts (Ayton, 2010; Gigerenzer, 1997; Ostrom, 2003; Simon, 1983; Sober and Wilson, 1998; Sutherland, 1992). For example, the take-the-best strategy (Gigerenzer, 1997) is a proposal that binary decisions (choosing one or the other) are made using as few cues as possible (cf. Simon, 1955). It is plausible that a "take-the-best" strategy is applicable to binary decisions in the social realm too (e.g., decide to interact with someone or not).

Sigmund (2010) delineated the mechanics of direct reciprocity in game theory terms (below described verbally, not mathematically). First, consider the Nash equilibrium (see also Binmore, 2005 and Ostrom, 2003) by imagining that player 1 can play two possible strategies — *e1* and *e2* — against player 2. This is a probability: if I play strategy *e1* this time, what is the probability that I will play the same strategy (*e1*) *next* time? It all depends on player 2, who (for example) has two possible strategies of her own: *f1* and *f2*. The Nash equilibrium is all about your "best response" to the other person's strategy. Suppose player 1 chooses *e1* and player 2 chooses *f1* — should player 1 stick to *e1* or switch to *e2*? If the payoff is higher by playing *e1* (instead of *e2*) in response to *f1*, then player 1 will likely keep on playing *e1* (his best response to *f1*). Remember, though, that the other's dyadic game *also* consists of two strategies. Player 2 might have her own Nash equilibrium — for example that *f1* is the best response to *e1*. Accordingly, the players can form an *equilibrium pair* and keep going in that same pattern which would prove beneficial for both. Suppose, however, the player 2 changes her strategy to *f2*. This might change the payoffs for player 1 and perhaps his new best response is now *e2* (or perhaps it stays the same). Thus, the game changes according to the behaviour on both sides of the dyad. Establishing equilibrium pairs through repetition is key to establishing a cooperative relationship within a dyad. Based on this, there have been many models of dyadic cooperation in the literature (see Binmore, 2005; Bowles and Gintis, 2011; Dugatkin, 1997, Sigmund, 2010). Games, of course, can have multiple equilibria, not just two (Binmore, 2005). Furthermore, there is huge variability of human behaviours, meaning that Nash equilibria are often reached at the group rather than individual level (Ostrom, 2003). As Alexander (1987) wrote, successful sociality is about "flexible strategizing" (p. 9). Such games, as described above (and laboratory experiments designed to test them), cannot begin to capture all of the messy complexity of real life (Binmore, 2005; Ostrom, 2003): nonetheless, such games are useful tools for understanding the principles that explain behaviour.

Imagine a different scenario where an organism is not only two, but many — living in a finite population where individuals might have three possible dyadic strategies (Sigmund, 2010): (1) always cooperate, (2) always defect (i.e., never give anything), (3) be choosy and do tit-for-tat (cooperate when meeting a co-operator, defect when meeting a defector). There have been many agent-based computer simulations where individual agents are programmed to use only *one* strategy each (e.g., Nowak and Sigmund, 1998). A typical such simulation comprises a series of iterations (one generation after another) and, after the program starts running, we assess the simulation by looking at how the proportions of types (cooperate/defect/choosy) alter over time. There are many questions to ask here. What type of agent will succeed in such a simulation? Will the population be overtaken by defectors? Will the co-operators prevail? Or, is choosiness the only path that allows cooperation to flourish? You can look at a population and

see how the percentages of each changes over time (e.g., see Bowles and Gintis, 2011; Nowak and Sigmund, 1998; Sigmund, 2010). For example, if the simulation begins with 100% co-operators, then it is fairly easy for defectors (who might appear suddenly due to mutation) to take over; but the defectors cannot dominate the population if discriminators are present. It all depends on proportions (Sigmund, 2010; Sober and Wilson, 1998): the initial mix of types, and how the interactions cause some groups to benefit over others (e.g., co-operators diminish in the population because they are giving too much away and not getting anything in return). The idea is that those who have successful strategies replicate (produce offspring using the same strategy) while those using unsuccessful strategies head towards extinction (if you don't get enough favours, you don't live to replicate). Therefore, strategies are said to "evolve" (increasing, decreasing, or staying about the same across iterations). How does a scientist decide the outcome of this complex mix? The possible outcomes can be derived from the *replicator equation* (described mathematically in Bowles and Gintis, 2011 and Sigmund, 2010). The replicator equation helps to predict how quickly a particular strategy will grow within this finite population, and the answer is that "a strategy *ei* will spread or dwindle depending on whether it does better or worse than average" (Sigmund, 2010, p. 31). The equation takes the average payoff to a particular strategy (e.g., a co-operator) and subtracts from that the average payoff of all individuals using all strategies (e.g., co-operators, defectors, and discriminators). Who tends to win, then? Generally, it depends on the numbers (how much percentage of each exists in the first place), but the discriminating strategies often win out. Defectors lose out when discriminators notice they are defecting; co-operators lose out because they are not choosy — but they can flourish when there are few or no defectors around (Bowles and Gintis, 2011; Dugatkin, 1997; Sigmund, 2010). Sometimes one can even get *rock-paper-scissor* dynamics (Sigmund, 2010): the proportions of each strategy oscillate, whereupon every strategy takes turns in being the most common.

A primary lesson here is that it pays to be choosy (Sober and Wilson, 1998). Referring to the environment of ancestral humans, Bowles and Gintis (2011) wrote that "those who failed to distinguish between long-term or short-term or one-shot interactions would be at a significant fitness disadvantage" (p. 96). In other words, treating everybody as a trusted friend will not necessarily benefit you. Given this risk, there must be some underlying principle that explains why people habitually act pro-socially, even to strangers.

Reputation and Indirect Reciprocity

Add a third person to a dyadic interaction and a triad emerges (Faust, 2007). When a three-way interaction occurs, it is nearly impossible for each actor to engage in precisely 33% of the interaction. There is inevitably some imbalance, with two of the actors more deeply involved than the third (cf. chimpanzee grooming

patterns described by Russell, 2007). Sometimes the third actor is not directly involved at all, but is merely watching the interaction. The proportion of non-involved individuals will increase further as the group size increases. This situation (being a triad or higher) sets the scene for indirect reciprocity. As Alexander (1987, pp. 94–95) described it:

> I regard indirect reciprocity as a consequence of direct reciprocity occurring in the presence of interested audiences — groups of individuals who continually evaluate the members of their society as possible future interactants from whom they would like to gain more than they lose (this outcome, of course, can be mutual).

In real life, cooperation is multidirectional. This is true throughout nature: all the way from the level of RNA hypercycles to that of human teamwork (see Bourke, 2011). Many models of cooperation have focused on the dyad (e.g., Bowles and Gintis, 2011; Dugatkin, 1997; Ostrom, 2003; Trivers, 1971). These dyadic models are predicated on the phenomenon of *trust* (see Kohn, 2009) — as well as using your knowledge of past behaviour to guide your current behaviour (for reviews and discussions, see Alexander, 1987; Dugatkin, 1997; McElreath, Clutton–Brock, Fehr, Fessler, Hagen, Hammerstein et al., 2003). Evidence from human studies shows that face-to-face contact is very important in establishing trust (Ostrom, 2003); but face-to-face contact cannot always happen. As a population grows larger, the probability of repeated interactions is reduced (because a person encounters strangers more and familiars less). In this case, being choosy is an essential strategy, because an indiscriminately generous person in a mixed population will always end up with a lower payoff than defectors (see Sober and Wilson, 1998, pp. 19–23). If learning whether to trust someone depended solely on direct encounters, then helpers are vulnerable to defection when helping a stranger the first time (Pollock and Dugatkin, 1992). This is a problem that can be by-passed if the helper has prior knowledge of how the potential recipient behaved in the past towards others. Reputation is useful here. It is that knowledge source.

Alexander (1977, 1987) proposed that indirect reciprocity ("A observes B help C, therefore A helps B") — a system that rewards the generous and punishes the selfish — is a defining mechanism of human moral systems (also see Binmore, 2005; Bowles and Gintis, 2011; Dugatkin, 1997; McElreath et al., 2003; Nowak and Highfield, 2011; Sigmund, 2010). Indirect reciprocity can occur in other forms too, such as "A helps B, B helps C, C helps A" (Alexander, 1987, p. 81), a form which will not be covered here. Another name for indirect reciprocity is "vicarious reciprocity" (Sigmund, 2010). The population-level benefit of indirect reciprocity might be simply summarised by saying that "everyone may gain when social beneficence is prevalent" (Alexander, 1987, p. 210; cf. Kohn, 2009). However, indirect reciprocity is *not* a synonym for "generalized reciprocity" (Alexander, 1987, p. 85). According to indirect reciprocity, when people are good, the strategy is

ultimately self-serving despite the up-front costs (Alexander, 1987). The benefits of indirect reciprocity towards a well-regarded individual can manifest in at least three ways: (1) direct compensation from all or part of a group (e.g., when someone is deemed a hero), (2) more opportunities to engage in fruitful interaction due to being approached by third parties who witnessed the generosity, and (3) the generosity ultimately benefits the group to which the generous person is a part — and perhaps even benefiting that person's own descendants (Alexander, 1987, p. 94).

Importantly, the terms indirect *reputation* and indirect *reciprocity* should not be confused. The former refers to an information source and the latter refers to the moral/social system that is enabled by the information source (Alexander, 1987; Nowak and Sigmund, 1998). Reputation is construable as one component of an interacting system where repeated social dilemmas (Dawes, 1980; Ostrom, 2003) are worked through when reputation feeds into trust, which feeds into the probability of reciprocity, leading hopefully to the best collective outcome possible (see Ostrom, 2003, pp. 49–61). Indirect reciprocity requires that group members monitor each other's reputations, ideally creating conditions where generous individuals prosper and selfish individuals suffer (Alexander, 1987; McElreath et al., 2003; Nowak and Sigmund, 1998; Pollock and Dugatkin, 1992). Over the years, a series of agent-based computer simulations have been developed to explore this possibility using an image scoring paradigm (Brandt and Sigmund, 2005; Nowak and Sigmund, 1998; reviewed in McElreath et al., 2003). "Image score" is a numerical measure of how generous an individual person has been to others. The aim of these simulations was to explore the conditions under which image scoring individuals (those who preferentially give rewards to those with sufficiently high image scores) would dominate a population that also consists of defectors (never help anyone) and unconditional givers (help others indiscriminately). The main conclusion from these models is that helping is an evolutionarily stable strategy (the population resists being overwhelmed by defectors; see Parker and Smith, 1990) only if the majority of the population consists of *strict* image scorers (Brandt and Sigmund, 2005; Nowak and Sigmund, 1998). The models inspired a series of real-life human experiments which showed that people actually do spontaneously consider reputation when deciding whom to reward; and that they behave more cooperatively in order to preserve their good reputations (Engelmann and Fischbacher, 2009; Wedekind and Milinski, 2000, etc.). Important to note, however, is that indirect reciprocity is not the only mechanism for preventing defections. Many large-scale human endeavours come about through institutionalisation — entailing the creation of formal organisations where things are put in writing and mechanisms designed to put principles above individual proclivities are in place (Alexander, 1987; Fehr and Gächter, 2000; Urpelainen, 2011). For example, think about the massive amount of planning and cooperation needed to successfully operate a highly complex entity such as London Heathrow Airport (Wicks, 2014): multiple levels of organisation are needed to manage

more than 76,000 people — each with a unique set of skills and stipulations — in their respective roles all geared towards the simply-stated (yet highly tricky to coordinate) goal of managing airplane arrivals and departures (an average of 1400 per day). Heathrow is a conspicuous example — but there are countless other types of organisations, large and small, that would be difficult or impossible to run without coordinated (and often highly regimented) action between strangers. Institutions often save us the trouble at needing to gather social information helping us decide with whom to work. In a place like Heathrow Airport, one does not usually need to know the reputations of those with whom one cooperates in order to get a plane to fly: people know each other's roles by default (baggage handler, pilot, etc.) and can therefore successfully collaborate with complete strangers constantly. Defecting is minimised through a set of rules and punishments (see Fehr and Gächter, 2000 for a review of reciprocity and punishment in the workplace; cf. Sober and Wilson, 1998). We might consider Heathrow Airport as a highly codified, almost reputation-irrelevant zone. This is one extreme on a spectrum of social situations. Another extreme is a setting consisting of familiars only: the kin, the friends, the neighbours, etc. This is where an abundance of information about the people one can interact with lies: not only that of an individual person, but the relationships between those individuals. Non-human animals, of course, usually exist (in the natural world) only amongst familiars (Hinde, 1976). It is we humans who cast the net wider.

Let us think again about the *replicator dynamics*, this time for indirect reciprocity (Bowles and Gintis, 2011; Sigmund, 2010), referring (as before) to the simulated population consisting of co-operator, defector, and discriminator. Now, the discriminator (the choosy one) needs to rely on the image score, which in a computer simulation can be as simple as 0/1 (known to be either generous in the previous round or not). The replicator dynamics equation here needs to incorporate the payoff for the reciprocator, who gives out a benefit only if the recipient has a good reputation, or when no information is available (in other words, reciprocators cooperate except when encountering a bad reputation). How can a discriminating strategy evolve in this setting in order to produce a simulation where cooperation is dominant? This depends, first of all, on whether reputations are knowable. As Sigmund (2010) wrote, "if the probability… to know the co-players' past is too small (i.e., if there is not much scope for reputation), then cooperation cannot evolve.… [A] cooperative population consisting of these two types of altruists (some conditional and some not) exists, if the average level of information within the population is sufficiently high" (p. 86). Once reputation becomes possible, then it all depends on the numbers: what percentage of the population is occupied by either co-operators, defectors, or discriminators. Obviously, too high a percentage of defectors will not allow cooperation to flourish — and too high a percentage of "gullible" co-operators will simply allow the defectors to take over unbridled. What is needed is a high-enough proportion of

discriminators to be the bulwark against the takeover by the selfish. A cooperative population can exist, even if there are many defectors and many gullibles, as long as the largest group happens to consist of discriminators (see Sigmund, 2010 for the mathematical treatment and implications of varying parameters). Now, let us think again about the knowability of reputation. Thinking across all animals, this depends very much on cognitive ability. Examples like Heathrow irresistibly remind us of insect societies, such as leafcutter ants (Nowak and Highfield, 2011) where success is implemented by seven different castes (anatomically differentiated) to carry out specialised tasks (within a world of chemical signalling). Clearly, complex cooperation arises non- or minimally cognitively across all facets of life (Bourke, 2011). This is why it is important to clarify the issue of when and why indirect reciprocity needs deliberation and when it does not.

The Cognitive Substrate of Indirect Reciprocity

In encountering animal studies, the temptation is often to infer human-style cognition. It is obvious that animals collect information (McGregor, 2005). As an information source, indirect reputation will flow ubiquitously from any animal communication network where it is possible to eavesdrop without being directly involved (Markl, 1985; McGregor, 2005).[1] The question, if we are thinking across the animal kingdom, is: What depth of processing occurs in animals living within these communication networks (Russell, 2007)? Evolutionary and psychological explanations of cooperative behaviour are interrelated (Sober and Wilson, 1998, pp. 203–205). It is useful to think of biological explanation the way Tinbergen (1963) delineated, in which every biological explanation can be construed in four ways: (1) phylogenetic (how it evolved), (2) ontogenetic (how it develops), (3) functional (why it evolved), and (4) proximate (the actual mechanism that enables it). In the animal kingdom, all examples of indirect reciprocity will have a functional explanation (number 3 above). Kolm's (2000) prescription for reciprocity can be applied to indirect reciprocity: thinking separately about allocation (the actual budget of given and received items) and process (the mechanism that enabled it). The questions of interest to psychologists tend to be those of process, that is, the proximate mechanisms (usually favouring cognitive explanations, with a special bias towards assuming conscious awareness). It is often difficult to write about non-sentient evolutionary processes in a way that does not sound like one is writing about characters in a play. This is why it is important to reiterate Tinbergen's "four whys." The issue of what is happening in the animal's mind when it engages in indirect reciprocity is a proximate-level description. At its simplest, indirect reciprocity is describable in terms of a dyadic interaction

[1]Here, I use the word "eavesdrop" to refer to any modality, whether it is from sight, sound, chemical senses, or other means.

(e.g., "A → B") that would not have occurred unless the actor had been primed by having observed an earlier interaction involving the recipient (e.g., "A had observed B → C"). Of course, this structure can also describe punishment (e.g., "A attacked B, because earlier B attacked C"). Three conditions are necessary for (positive) indirect reciprocity to occur: (1) favours occur in a setting observable by third parties, (2) the third party is motivated to reward the rewarder, and (3) the third party is influenced by indirect reputation.

The first condition, observability of behaviour, is the prerequisite for all phenomena involving reputationally based cooperative behaviour. Among humans, behaviour might be observed second- or third-hand through verbal gossip, whereas pre-linguistic animals are limited to direct observation. A wide variety of species have evolved observational skills that effortlessly detect cues and signals emitted from others regardless of where the emitter was aiming (it is possible, of course, that a signal is emitted without an intended direction). In this context, a field of *public information* evolves in the mind's eye of the species — the populations of which are now able to interpret "inadvertent social information" (Danchin, Giraldeau, Valone, and Wagner, 2004). The second condition concerns the proximate motivation of the third party in repaying the favour on behalf of the recipient. There are four ways to partition this, as explained below.

1. *There is no motivation (on a cognitive level).* Above, indirect reciprocity was described in its simplest form: an interaction occurring as a consequence of the actor being primed by a prior interaction involving the recipient and a third party. Defined this way, indirect reciprocity can be identified anywhere the above causality is established, regardless of the level of cognition of the actors. Some examples are found among cleaner fish and their "clients" (Bshary, 2002) ["client" refers to the recipient of prosocial behavior, e.g., other species of fish]. Here, clients observe the interactions of cleaner fish towards the third parties; those cleaners who defect (eat living flesh in addition to the dead flesh that they are supposed to be cleaning off) are avoided by the clients more than cleaners who do not cheat (Bshary, 2002). Indirect reciprocity is likely restricted here to a functional rather than an explanation. How much the fish actually understands the third-party interactions is open to debate, but the point shown by Bshary is that such events can be identified in cases where sophisticated cognitive abilities are unproven (it is possible, of course, that fish are cleverer than we think — but we are safest for now in assuming that indirect reciprocity is happening with minimal cognition in this class of animal).

2. *There is a selfish motivation.* Alexander (1987) suggested that individuals may reward the rewarder simply as a by-product of their desire to interact with someone known to be cooperative. If cleaner fish clients (Bshary, 2002) were human-style conscious beings, then their motivations might be regarded as selfish: they reward non-cheating cleaner fish by offering themselves, in the process rewarding both themselves (being cleaned) and the cleaner (who obtains food).

Traditional economic theories predict humans to behave this way, but empirical results show that humans in economic games behave selfishly only part of the time (Binmore, 2005; Schram, 2000). This implies that humans have something to gain by behaving pro-socially (see next point). Interestingly, chimpanzees seem more likely than humans to behave in a self-interested manner consistent with traditional economic theories (Jensen, Call, and Tomasello, 2007).

3. *There is a motivation to behave pro-socially "for its own sake."* Schram (2000) identified three reasons that humans might pursue a conscious pro-sociality. The first is the "warm glow of giving" where an individual cooperates because it feels good to do so (see also Binmore, 2005; Pradel and Fetchenhauer, 2010; Sober and Wilson, 1998, pp. 267–271). As mentioned above, such feelings could arise by association with past positive outcomes. This means that being generous (e.g., giving to a charity to help starving children) can actually be selfish: "If you were an egoist, you would help the starving, but your ultimate motive would be to make yourself feel good" (Sober and Wilson, 1998, p. 244). The second reason is fairness, where an individual cooperates on the contingency that the partner co-operates. This is the basis of a number of tit-for-tat cooperation models (Dugatkin, 1997; Trivers, 1971), which in humans involves knowing reciprocity norms (Binmore, 2005). Operating in this manner requires that an individual engage in mental score-keeping: keeping track of all past activities of one's trading partners (see also Call, 2002; Schino and Aureli, 2009). Theoretical models have introduced many variations on this general idea (e.g., contrite tit-for-tat; see Dugatkin, 1997). Among apes, mental score-keeping is unproven but possible (Call, 2002; Schino and Aureli, 2009). Whether apes can really recognise reciprocity norms is unclear (cf. Tomasello, Carpenter, Call, Behne, and Moll, 2005). The third reason identified by Schram (2000) was other-regarding: concern for the well-being of someone else (perhaps emotionally charged). All of the above reasons could apply to indirect reciprocity.

4. *There is a motivation based on consciously known cooperative gains.* This occurs when an individual understands collaborative behaviour (Schram, 2000): that a goal cannot be accomplished unless individuals work together (cf. Dawes, 1980; Dugatkin, 1997; Ostrom, 2003). This is not a difficult task if two individuals are standing side-by-side co-operating to achieve an immediate goal. Even chimpanzees can do this if trained, but only if the reward is forthcoming to both. If one chimpanzee does not receive an immediate observable reward, then that individual will probably not cooperate (Jensen, Hare, Call, and Tomasello, 2006). In the case of indirect reciprocity, this would require an explicit understanding concerning how indirect reciprocity maintains the overall pro-sociality of the group. In real-life terms, this might be seen among human societies where community spirit plays a salient role in people's lives, for example, in the Indonesian farming community that Schweizer (1989) studied, where the concept of neighbourly harmony was an influential factor in social and religious life. This prevented

families from adopting practices that maximised financial income at the expense of community members (cf. Dawes, 1980). Two examples were cited: (a) farming families chose labourer hiring practices that distributed wealth more evenly, and (b) families never eschewed their obligation to host a *slametan*, a religiously-based feast that was thought to spiritually benefit the whole community (despite the expense to the family). These examples represent a form of indirect reciprocity, because generous behaviour is rewarded (but not in a dyadic tit-for-tat manner). Although some of this pro-social behaviour is probably motivated by a fear of ostracism (Schweizer, 1989), everyone is aware of the relationship between one's own behaviour and the community's well being (cf. Urpelainen, 2011).

As illustrated, there are many paths to indirect reciprocity. It is a behavioural example of the evolutionary principle of functional equivalence, meaning that it is possible to identify a number of different proximate mechanisms which "all deliver roughly the same behaviors in the same circumstances" (Sober and Wilson, 1998, p. 206). Indirect reciprocity is a beneficial strategy in particular circumstances, and different animals have evolved different levels of necessary cognition. Among humans, each of the above levels of motivation should occur. Moreover, selfish and pro-social motivations can co-exist and intermingle (for example, "person A might have initially helped B for selfish reasons, had been rewarded, and then started to care for B's welfare without wanting a reward," cf. Sober and Wilson, 1998, pp. 217–222, 242–250, 319–321). Although we congratulate ourselves on being able to adopt the most cognitively sophisticated strategies in the animal kingdom, much of the time we are likely getting by with a minimum of cogitation (Newell and Shanks, 2014): indirect reciprocity may occur unplanned, be a by-product of a selfish motivation, or follow genuinely pro-social sentiments motivated either by a feel-good factor or from a self-aware intention to contribute to the common good (see Ostrom, 2003). Among non-humans, there is still much to learn about how this works. To address this question, the best approach is probably to emulate approaches equivalent to Byrne and Whiten (1997) when they searched for deception in the animal kingdom (see also Byrne, 2003): find as many instances as possible where deception occurs and only later start worrying about characterising the cognition (if any) that is involved. We can do the same for indirect reciprocity.

Thus, when indirect reciprocity occurs, premeditation may not be necessary. It should be useful to identify instances of indirect reciprocity in nature as a possible context where individuals are heeding each other's reputations. The evidence for indirect reciprocity among non-human primates is sketchy, but there are possible candidates in the literature. One example is the possible revenge system described by Aureli, Cozzolino, Cordisch, and Scucchi (1992) in macaques, where individuals would attack the family member of an aggressor. Another example is in a catalogue of "triplet interactions" by Mori (1983) on free-ranging chimpanzees: sequences of behaviours where a dyadic interaction was soon followed by a different

dyadic interaction involving one of the previous interactants (see Mori, 1983, Table 8). Twenty different types of triadic interaction were identified, two of which may qualify for indirect reciprocity. One interaction type was where "a first appeasing chimpanzee became the recipient in the second interaction" (Mori, 1983, p. 58, #4 in Table 8). The other one was a revenge incident (#9 in Table 8) similar to that reported by Aureli et al. (1992). Reputation probably plays a role in such interactions. For a researcher interested in reputational thinking, the key issue is whether an individual can compare relationships between self and other to relationships among third parties (Russell, 2007). At best, we can take a "bird's eye view" and imagine living within a complex social structure and be aware of our place in it. Most of the time, we are not doing this explicity.

Social Expertise

It would be valuable to apply cognitive–psychological models of expertise to the social and non-human sphere. I will start with what Markl (1985, p. 165) said about the recipient of animal signals:

> Of course, addressees are not sitting around in extra-evolutionary space offering [unmodifiable] releasing mechanisms just waiting to be manipulated; in fact, it is well known that there is hardly anything that can be more easily modified both by evolution and individual experience — where we call it focusing of attention — than reaction thresholds and response selectivity of releasing mechanisms or sensory-neural pattern recognition devices.

Not learning means not surviving. "Inflexibility or preprogramming would be the worst possible strategy in the face of conflicts of interest, competition, the importance of cooperation, and other aspects of sociality" (Alexander, 1987, p. 9). The history of interactions that leads to something we can call a "relationship" (Hinde, 1976) is also the story of successive learning experiences, gauging and re-gauging expectations (Markl, 1985). This might lead to something we can call expertise.

Social expertise is something that Humphrey (1976) compares to chess: a game played with a reactive partner, where competence depends on accumulated knowledge, ability to keep track of changeable circumstances depending on the opponent's behaviour, and planning ahead according to what others may do. The Machiavellian intelligence hypothesis (Byrne and Whiten, 1997) construed social expertise as a skilled manipulation of others for personal gain. This requires "mind reading" ability (Byrne and Whiten, 1997), which is generally the skill of visualising the point of view of another's perception and intentions, seeing how intervening variables alter such intentions, and being able to identify deceptive behaviours (cf. Tomasello et al., 2005). This view of expertise can be nested within a broader framework called the social brain hypothesis (Dunbar, 2003; Dunbar,

Gamble, and Gowlett, 2014), where function is emphasized: the burden of maintaining an optimal size of one's personal social network within larger social groups. According to this view, ecological factors put pressure on individuals to form large groups in order to enhance survival (Bourke, 2011; Dunbar, 1988), which in turn creates selective pressure for the evolution of increased size in the areas of the brain that facilitate social expertise (Dunbar, 2003). There appears to be a widely held opinion that the key skill of manipulation is an ability to anticipate another's behaviour based on a talent for mind reading (a.k.a. theory of mind or experience projection) [e.g., Byrne and Whiten, 1997; Dunbar, 2003; Humphrey, 1976]. The term "mind reading," of course, has been used to describe intention reading even in less cognitively advanced animals, referring to an animal's innate reactions to certain cues and signals (cf. Krebs and Dawkins, 1984). What makes mind reading "expert" is the knowledge base that the animal draws upon in order to behave proficiently. As Donald (2001) wrote, such mental feats "demand considerable memory, since each individual must have a 'slot' in the tracker's mind, which must be kept up to date" (p. 129). Monkeys and apes accumulate considerable knowledge about their conspecifics in at least three domains (Call, 2002): (1) information about how individuals behave, (2) the quality of their relationships with others, and (3) the quality of relationships among third parties. This type of knowledge base permits an individual to engage in social manipulation: where a manipulator induces a conspecific to behave in a certain way in order to accomplish a goal desired by the manipulator. For example, begging to receive food is a form of dyadic manipulation ("A induced B to do X") [Call, 2002]. A more complex skill is triadic manipulation (Call, 2002), where a manipulator induces a conspecific to behave a certain way towards a third party ("A induced B to induce C to do X"). Both dyadic and triadic manipulations are known as "social tool use," an expert skill that requires accumulated knowledge about the typical behaviour of others, along with some ability "to generate hypotheses about who interacts with whom, when and how" (Call, 2002, p. 178).

Even dyadic manipulation requires some form of indirect reputation — because the only way for A to learn the causal chain between a conspecific and a desired outcome is to observe how B behaves towards something in the environment, and to see how that behaviour leads to the outcome. What differentiates this from technical tool use (observing how a tool behaves towards the environment) is the fact that the social tool involves an animate being (Call, 2002), and hence there is a built-in source of uncertainty (whether B will behave the way anticipated). Triadic manipulation entails the same observational learning as dyadic manipulation, except that the causal chain now has two sources of uncertainty (whether either B and C will behave the way anticipated). How might an individual overcome such uncertainty in order to make the social manipulation work? Pure luck is obviously a factor (but one that likely underpays). What might facilitate higher payoffs is predictive ability, where individual A has learned — by experience — what behaviours to

expect from individuals B and C during the relevant events. To illustrate this advantage, consider the following hypothetical situation about food sharing.

Imagine that individual A observes that C is monopolising a local food source (i.e., can prevent others from getting it). A desires the food, but knows from past experience that C will not offer any food if approached. However, A knows two other things:

(1) if B approaches C, then C always gives food to B

(2) if A (self) approaches B, then B always gives food to A

Here, triadic manipulation can occur if A induces B to approach C for food. When B obtains the food and carries it away, this provides an opportunity for A to beg for food from B. Additionally, suppose that individuals D and E are also nearby, but that A knows three other facts:

(3) if D begs for food from C, then C will refuse

(4) if A begs for food from E, then E will refuse

(5) if E begs for food from C, then C will give food to E

Obviously, it is pointless for A to engage D or E in social manipulation because D cannot obtain food from C, and E won't give it to A (even if getting it from C). If A knows this, then A will approach B and nobody else. This might comprise a form of declarative memory (Anderson, 1983): learned factual knowledge stored as long-term memory traces (interconnections). Facts in the long-term memory can be organised into themes, whereupon a number of thematically related facts are interconnected. This is one basis for an expert memory. It might be useful here to refer to general theories of expertise — in order to illuminate mechanisms that might also apply in the social domain. On the topic of non-social human expertise, there is a long and rich history of psychological testing and theorising (de Groot and Gobet, 1996; Ericsson, Charness, Feltovich, and Hoffman, 2006; Gobet, Chassy, and Bilalić, 2011; Russell, 2011; Sternberg, 1997). Some primate researchers have made detailed comparisons between physical and social reasoning, in an attempt to delineate commonalities and contrasts in the required intellectual abilities between species (e.g., Call, 2002). The cognitive mechanisms of human expertise have been characterised in many different ways over the years, but there is general agreement that expert skill acquisition involves deliberate practice, learning a large number of relevant patterns, cultivating a long-term memory base where memory traces are flexibly accessed, and understanding how to respond appropriately to meaningful patterns (Anderson, 1983; Gobet et al., 2011; Russell, 2011; Sternberg, 1997). In the food-sharing example presented

above, individual A knew the reputations of individuals B, C, D, and E based on past experience. As mentioned earlier, typicality is key (cf. Emler, 1990). Another key issue in expertise is the amount of accumulated knowledge (Sternberg, 1997), and this is applicable in the social realm too.

Table 1

Third Party Knowledge of Possible Direct Benefits Based on Indirect Observation of Focal Animal's Current Behaviour to Someone Else

Knowledge Base (friend)
A watches what B (friend) is doing to F:
If B → √ → F, then B → √ → A 100% of time
If B → X → F, then B → √ → A 100% of time
Knowledge Base (semi-friend)
A watches what C (semi-friend) is doing to F:
If C → √ → F, then C → √ → A 100% of time
If C → X → F, then C → √ → A 70% of time and C → X → A 30% of time
Knowledge Base (non-friend)
A watches what D (non-friend) is doing to F:
If D → √ → F, then D → √ → A 100% of time
If D → X → F, then D → √ → A 20% of time and D → X → A 80% of time
Knowledge Base (enemy)
A watches what E (enemy) is doing to F:
If E → √ → F, then E → √ → A 100% of time
If E → X → F, then E → X → A 100% of time

Note: This is a kind of rudimentary classification system; here, the individual has a non-verbal "knowledge base" about each category of friend, semi-friend, non-friend, and enemy, which facilitates an appraisal of what is likely to happen. It implies knowledge about each of the individuals involved. Thus, you are less likely to approach an enemy for food because he will likely refuse you even if he has been seen feeding someone else. At one extreme, there is the friend (100% chance of feeding you if he fed someone else). At the other extreme, there is the enemy (0% chance). There are also two other situations (semi-friend, non-friend) which we can regard as representing two points along a continuum between the extremes. The X refers to hostile behaviour (e.g., attack). The check mark (√) refers to friendly behaviour (e.g., feed). Arrows indicate the direction of these behaviours. For example, B → √ → F means that individual B is friendly to individual F; and B → X → F means that B is hostile to F.

Employing an "if–then" syntax in social reasoning (de Waal, 2003), Table 1 (above) is a conjectural framework showing how a person (possibly also an animal) might employ this syntax when facing others who represent four different grades of social relationship: (1) a friend, (2) a semi-friend, (3) a non-friend, and (4) an enemy (all numbers are notional). These grades do not presume formal labels in the mind (cf. Spears, 2011). What these labels represent will, however, somehow have real-life denotation. Specifically, Table 1 shows a scenario where an animal begins to recognise differences in direct experience that *correlate* with observed third-party interactions. Observer A will know how every individual (friend, semi-friend, non-friend, enemy) is likely to behave towards A, after observing how these individuals behaved towards others. In all cases, in the table, friendliness begets friendliness. Differences arise in what happens after the conspecifics are observed being hostile towards others. The friend is the easiest to comprehend and trust, because the behaviour is friendly to A 100% of the time. The second easiest to comprehend is the enemy, who, if hostile to others, is hostile to A. The behaviour of the non-friend is less predictable: if hostile to others, the non-friend is usually — but not always — hostile to A. In the case of the semi-friend, there is also unpredictability: if hostile to others, the semi-friend is usually — but not always — friendly to A. The rate of hostility is low (only 30%), which means that the observer should regard friendliness as the default expectation. The if–then syntax is the basis upon which an observer develops an understanding about the correlations between direct reputational experience (e.g., how B behaved towards me) and indirect reputation (e.g., how B behaved towards others). The knowledge of this correlation is an impression inside the animal's mind, established during a personal history between the observer and the other animal (information by direct reputation), and intuitively cross-checked against the other's interaction with others (information by indirect reputation).

Thinking of conspecific behaviour in a probabilistic manner (as above) is useful because some kinds of information are important for survival, such as avoiding attack: hostile behaviour necessitates that the observer be vigilant. As Dunbar (1988) wrote, the "amount of visual monitoring that an animal does is primarily a function of its nervousness, and reflects the animal's need to keep track of the movements of the more dominant individuals in order to avoid being attacked unawares" (p. 115). Situations like this are where it is advantageous to have an additional channel of information (e.g., indirect reputation) in addition to personal encounter (direct reputation). As Sober and Wilson (1998) wrote: "two sources of evidence are better than one, as far as reliability is concerned" (p. 307). When facing the conspecifics, as presented in Table 1, the observer would be vigilant when facing the enemy and non-friend (because they might attack); and non-vigilant when facing the friend and semi-friend (because they are unlikely to attack). For everyone but the enemy, expectations are based on some degree of trust (cf. Kohn, 2009). It is beneficial for an observer

to be non-vigilant most of the time, because this frees the observer's attention to focus on other things (e.g., feeding) [Dunbar, 1988]. Accuracy of assessment would be valuable here, enabling the observer to know when to relax, and with whom to associate. However, small sample sizes could create misleading impressions. To know that the semi-friend is friendly 70% of the time (as per Table 1), the observer should perhaps witness at least ten occurrences during which friendliness occurred seven times. If the observer has witnessed only one occurrence and it was hostile, a misleading impression has been formed — making the observer unnecessarily vigilant in the semi-friend's presence (an example where a "larger sample size," i.e., more encounters, would be useful). It seems clear that the direct and indirect experience would have unequal influence on that social impression. There is surely what Sober and Wilson (1998) call "D/I asymmetry": the direct (D) experience will likely be more reliable than indirect (I) experience. In gaining indirect knowledge, you may not have seen all of the relevant events that make an accurate impression; furthermore, the occurences are towards other people (not you). We can also make a note about the third form, gossip: it is a cheap form of information (cf. Smith and Harper, 2004) for a couple of reasons. Firstly, it is the most subject to distortion (you did not observe the events yourself, but learned about them through at least one other person's cognitive filters). Secondly, gossip is much easier to fake (because people lie) than information gained from observing behaviour directly, in particular when honest (difficult-to-fake) signals are being displayed (e.g., visual cues of health). So, comparing the different "channels," direct (D), indirect (I), and reported (R) reputation, in terms of value (e.g., reliability and accuracy), it is plausible that $D > I > R$.

Also, we should remember that our social impressions will be riddled with inaccuracies — a social version of the inaccuracies uncovered in decision theory (cf. Ayton, 2010). Furthermore, the social impression is likely highly distorted by emotional processing. As Schino and Aureli (2009) argue, cooperative behaviour amongst animals is likely mediated by a kind of "emotional bookkeeping" (rather than a rational and cognitive bookkeeping). This is likely in humans too (McElreath et al., 2003), because people generally are not rational actors maximizing their benefits without emotion (Gigerenzer, 1997; Simon, 1955, 1983; Sober and Wilson, 1998; Sutherland, 1992). Binmore (2005) proffered that emotion "evolved to help police primeval social contracts, and they remain useful to us for this purpose" (p. 83). A behaviourist interpretation of emotional bookkeeping is that organisms are motivated by emotional rewards and punishments that get associated with specific interactions with particular individuals (see discussion in Sober and Wilson, 1998, pp. 256–260). To me, this sounds like the basis for acknowledging at least a rudimentary form of social expertise in animals (cf. Helton, 2005), applicable to the concepts of reputation and reciprocity, direct and indirect.

References

Alexander, R. D. (1977). Natural selection and the analysis of human sociality. In C. E. Goulden (Ed.), *Changing scenes in the natural sciences, 1776–1976* (pp. 283–337). Philadelphia: Academy of Natural Sciences.

Alexander, R. D. (1987). *The biology of moral systems*. New York: Aldine de Gruyter.

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, Massachusetts: Harvard University Press.

Aureli, F., Cozzolino, R., Cordisch, C., and Scucchi, S. (1992). Kin-oriented redirection among Japanese macaques: An expression of a revenge system? *Animal Behaviour*, *44*, 283–291.

Ayton, P. (2010). Judgement and decision making. In H. Kaye (Ed.), *Cognitive psychology* (second edition, pp. 417–453). Milton Keynes: The Open University.

Barnard, C. J., and Sibly, R. M. (1981). Producers and scroungers: A general model and its application to feeding flocks of house sparrows. *Animal Behaviour*, *29*, 543–550.

Binmore, K. (2005). *Natural justice*. Oxford: Oxford University Press.

Bourke, A. F. G. (2011). *Principles of social evolution*. Oxford: Oxford University Press.

Bowles, S., and Gintis, H. (2011). *A cooperative species: Human reciprocity and its evolution*. Princeton: Princeton University Press.

Brandt, H., and Sigmund, K. (2005). Indirect reciprocity, image scoring, and moral hazard. *Proceedings of the National Academy of the United States of America*, *102*, 2666–2670.

Bshary, R. (2002). Biting cleaner fish use altruism to deceive image-scoring client reef fish. *Proceedings of the Royal Society of London Series B: Biological Sciences*, *269*, 2087–2093.

Byrne, R. (2003). Tracing the evolutionary path of cognition. In M. Brüne, H. Ribbert, and W. Schiefenhövel (Eds.), *The social brain: Evolution and pathology* (pp. 43–60). Chichester: John Wiley & Sons.

Byrne, R., and Whiten, A. (Eds.). (1997). *Machiavellian intelligence II: Extensions and evaluations*. Cambridge: Cambridge University Press.

Call, J. (2002). Social knowledge and social manipulation in monkeys and apes. In C. S. Harcourt and B. R. Sherwood (Eds.), *New perspectives in primate evolution and behaviour* (pp. 173–196). Otley, West Yorkshire: Westbury.

Carmeli, A., and Tishler, A. (2005). Perceived organizational reputation and organizational performance: An empirical investigation of industrial enterprises. *Corporate Reputation Review*, *8*, 13–30.

Chekhov, A. (1921). A slander [C. Garnett, Trans.]. In *The tales of Chekhov, volume X: The horse stealers and other stories* (pp. 221–226). New York: The Macmillan Company (originally published 1883).

Clutton–Brock, T. (2009). Cooperation between non-kin in animal societies. *Nature*, *462*, 51–57.

Danchin, E., Giraldeau, L.-A., Valone, T. J., and Wagner, R. H. (2004). Public information: From nosy neighbours to cultural evolution. *Science*, *305*, 487–491.

Dawes, R. M. (1980). Social dilemmas. *Annual Review of Psychology*, *31*, 169–193.

de Groot, A., and Gobet, F. (1996). *Perception and memory in chess. Heuristics of the professional eye.* Assen, Netherlands: Van Gorcum.

de Waal, F. B. M. (2003). Social syntax: The if–then structure of social problem solving. In F. B. M. de Waal (Ed.), *Animal social complexity* (pp. 230–248). Cambridge, Massachusetts: Harvard University Press.

Donald, M. (2001). *A mind so rare: The evolution of human consciousness*. New York: W. W. Norton & Company.

Dugatkin, L. A. (1997). *Cooperation among animals: An evolutionary perspective*. New York: Oxford University Press.

Dunbar, R. I. M. (1988). *Primate social systems*. London: Croon Helm.

Dunbar, R. I. M. (2003) The social brain: Mind, language, and society in evolutionary perspective. *Annual Review of Anthropology*, *33*, 163–181.

Dunbar, R. I. M. (2004). Gossip in evolutionary perspective. *Review of General Psychology*, *8*, 100–110.

Dunbar, R., Gamble, C., and Gowlett, J. (2014). *Thinking big: How the evolution of social life shaped the human mind*. New York: Thames & Hudson.

Emler, N. (1990). A social psychology of reputation. *European Review of Social Psychology*, *1*, 172–193.

Emler, N. (1994). Gossip, reputation, and social adaptation. In R. F. Goodman and A. Ben-Ze'ev (Eds.), *Good gossip* (pp. 117–138). Lawrence, Kansas: University Press of Kansas.

Engelmann, D., and Fischbacher, U. (2009). Indirect reciprocity and strategic reputation building in an experimental helping game. *Games and Economic Behavior*, *67*, 399–407.
Ericsson, K.A., Charness, N., Feltovich, P. J., and Hoffman, R. R. (Eds). (2006). *The Cambridge handbook of expertise and expert performance*. Cambridge: Cambridge University Press.
Faust, K. (2007). Very local structure in social networks. *Sociological Methodology*, *37*, 209–256.
Fehr, E., and Gächter, S. (2000). Fairness and retaliation: The economics of reciprocity. *Journal of Economic Perspectives*, *14*, 159–181.
Gigerenzer, G. (1997). Bounded rationality: Models of fast and frugal inference. *Swiss Journal of Economics and Statistics*, *133*, 201–218.
Gobet, F., Chassy, P., and Bilalić, M. (2011). *Foundations of cognitive psychology*. London: McGraw–Hill.
Greif, A. (1989). Reputations and coalitions in medieval trade: Evidence on the Maghribi traders. *The Journal of Economic History*, *69*, 857–882.
Guttman, N. (1953). Operant conditioning, extinction, and periodic reinforcement in relation to concentration of sucrose used as reinforcing agent. *Journal of Experimental Psychology*, *46*, 213–224.
Hamilton, W. D. (1975). Innate social aptitudes of man: An approach from evolutionary genetics. In R. Fox (Ed.), *Biosocial anthropology* (pp. 133–153). London: Malaby Press.
Helton, W. S. (2005). Animal expertise, conscious or not. *Animal Cognition*, *8*, 67–74.
Hinde, R. A. (1976). Interactions, relationships, and social structure. *Man*, *11*, 1–17.
Humphrey, N. K. (1976). The social function of intellect. In P. G. Bateson and R. A. Hinde (Eds.), *Growing points in ethology* (pp. 303–317). Cambridge: Cambridge University Press.
Jensen, K., Call, J., and Tomasello, M. (2007). Chimpanzees are rational maximizers in an ultimatum game. *Science*, *318*, 107–109.
Jensen, K., Hare, B., Call, J., and Tomasello, M. (2006). What's in it for me? Self-regard precludes altruism and spite in chimpanzees. *Proceedings of the Royal Society of London Series B: Biological Sciences*, *273*, 1013–1021.
Kohn, M. (2009). *Trust: Self-interest and the common good*. Oxford: Oxford University Press.
Kolm, S.-C. (2000). The theory of reciprocity. In L.-A. Gérard–Varet, S.-C. Kolm, and J. Mercier Ythier (Eds.), *The economics of reciprocity, giving and altruism* (pp. 115–141). Basingstoke: Palgrave Macmillan.
Krebs, J. R., and Dawkins, R. (1984). Animal signals: Mind reading and manipulation. In J. R. Krebs and N. B. Davies (Eds.), *Behavioural ecology: An evolutionary approach* (second edition, pp. 380–402). Oxford: Blackwell.
Markl, H. (1985). Manipulation, modulation, information, cognition: Some of the riddles of communication. In B. Hölldobler and M. Lindauer (Eds.), *Experimental behavioral ecology and sociobiology* (pp. 163–194). Sunderland, Massachusetts: Sinauer.
McElreath, R., Clutton–Brock, T. H., Fehr, E., Fessler, D. M. T., Hagen, E. H., Hammerstein, P., Kosfeld, M., Milinski, M., Silk, J. B., Tooby, J., and Wilson, M. I. (2003). Group report: The role of cognition and emotion in cooperation. In P. Hammerstein (Ed.), *Genetic and cultural evolution of cooperation* (pp. 125–152). London: The MIT Press.
McGarty, C. (2002). Stereotype formation as category formation. In C. McGarty, V. Y. Yzerbyt, and R. Spears (Eds.), *Stereotypes as explanations: The formation of meaningful beliefs about social groups* (pp. 16–37). Cambridge: Cambridge University Press.
McGregor, P. K. (2005). *Animal communication networks*. Cambridge: Cambridge University Press.
Mori, A. (1983). An ethological study on chimpanzees at the artificial feeding place in the Mahale mountains, Tanzania — with special reference to the booming situation. *Primates*, *23*, 45–65.
Newell, B. R., and Shanks, D. R. (2014). Unconscious influences on decision making: A criticial review. *Behavioral and Brain Sciences*, *37*, 1–61.
Nowak, M., and Highfield, R. (2011). *Super cooperators*. Edinburgh: Canongate.
Nowak, M. A., and Sigmund, K. (1998). Evolution of indirect reciprocity by image scoring. *Nature*, *393*, 573–577.
Obreiter, P., Fähnrich, S., Jens, N., and Gianluca, M. (2005). How social structure improves distributed reputation systems: Three hypotheses. *Lecture Notes in Computer Science*, *3601*, 229–236.
Ostrom, E. (2003). Toward a behavioral theory linking trust, reciprocity, and reputation. In E. Ostrom and J. Walker (Eds.), *Trust and reciprocity: Interdisciplinary lessons from experimental research* (pp. 19–79). New York: Russell Sage.

Parker, G. A., and Smith, J. M. (1990). Optimality theory in evolutionary biology. *Nature*, *348*, 27–33.
Pollock, G., and Dugatkin, L. A. (1992). Reciprocity and the emergence of reputation. *Journal of Theoretical Biology*, *159*, 25–37.
Pradel, J., and Fetchenhauer, D. (2010). Why most theories get it wrong: Altruistic intentions as an explanation of the evolution of genuine altruism. In U. J. Frey, C. Störmer, and K. P. Willführ (Eds.), *Homo novus – A human without illusions* (pp. 79–92). Berlin: Springer Verlag.
Russell, Y. I. (2007). *Reputations and polyadic interactions among great apes.* Doctoral dissertation, School of Biological Sciences, University of Liverpool.
Russell, Y. I. (2011). Prehistoric stone tools, chess expertise, and cognitive evolution: An experiment about recognizing features in flint debitage. *Journal of Evolutionary Psychology*, *9*, 249–269.
Russell, Y. I., Call, J., and Dunbar, R. I. M. (2008). Image scoring in great apes. *Behavioural Processes*, *78*, 108–111.
Russell, Y. I., and Phelps, S. (2013). How do you measure pleasure? A discussion about intrinsic costs and benefits in primate allogrooming. *Biology and Philosophy*, *28*, 1005–1020.
Schino, G., and Aureli, F. (2009). Reciprocal altruism in primates: Partner choice, cognition, and emotions. *Advances in the Study of Behaviour*, *39*, 45–69.
Schram, A. (2000). Sorting out the seeking: The economics of individual motivations. *Public Choice*, *103*, 231–258.
Schweizer, T. (1989). Economic individualism and the community spirit: Divergent orientation patterns of Javanese villagers in rice production and the ritual sphere. *Modern Asian Studies*, *23*, 277–312.
Sigmund, K. (2010). *The calculus of selfishness*. Princeton: Princeton University Press.
Simon, H. A. (1955). A behavioural model of rational choice. *Quarterly Journal of Economics*, *69*, 99–118.
Simon, H. A. (1983). *Reason in human affairs.* Stanford: Stanford University Press.
Smith, J. M., and Harper, D. (2004). *Animal signals*. Oxford: Oxford University Press.
Sober, E., and Wilson, D. S. (1998). *Unto others: The evolution and psychology of unselfish behavior.* Cambridge: Harvard University Press.
Spears, R. (2011). Group identities: The social identity perspective. In S. J. Schwartz, K. Luyckx, and V. L. Vignoles (Eds.)., *Handbook of identity theory and research* (pp. 201–224). New York: Springer.
Spencer, R., Russell, Y. I., Dickins, B. J. A., and Dickins, T. E. (2017). Kleptoparisitism in gulls *Laridae* at an urban and a coastal foraging environment: An assessment of ecological predictors. *Bird Study*, *64*, 12–19.
Sternberg, R. J. (1997). Cognitive conceptions of expertise. In P. Feltovich, K. M. Ford, and R. R. Hoffman (Eds.), *Expertise in context: Human and machine* (pp. 149–162). Menlo Park, California: AAAI/MIT Press.
Sutherland, S. (1992). *Irrationality*. London: Pinter & Martin.
Tinbergen, N. (1963). On aims and methods of ethology. *Zeitschrift für Tierpsychologie*, *20*, 410–433.
Tomasello, M., Carpenter, M., Call, J., Behne, T., and Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, *28*, 675–691.
Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology*, *46*, 35–57.
Urpelainen, J. (2011). The origins of social institutions. *Journal of Theoretical Politics*, *23*, 215–240.
Wedekind, C., and Milinski, M. (2000). Cooperation through image scoring in humans. *Science*, *288*, 850–852.
Wicks, R. (2014). *Heathrow airport*. Yeovil, Somerset: Haynes.