PHILOSOPHICAL TRANSACTIONS B

royalsocietypublishing.org/journal/rstb

Review

Cite this article: Santos FP, Pacheco JM, Santos FC. 2021 The complexity of human cooperation under indirect reciprocity. *Phil. Trans. R. Soc. B* **376**: 20200291. https://doi.org/10.1098/rstb.2020.0291

Accepted: 12 April 2021

One contribution of 20 to a theme issue 'The language of cooperation: reputation and honest signalling'.

Subject Areas:
behaviour, cognition, evolution

Keywords:
cooperation, reputation, indirect reciprocity, complexity, evolutionary dynamics, evolutionary game theory

Authors for correspondence:
Fernando P. Santos
e-mail: fpsantos@princeton.edu; f.p.santos@uva.nl
Jorge M. Pacheco
e-mail: jmpacheco@math.uminho.pt
Francisco C. Santos
e-mail: franciscocsantos@tecnico.ulisboa.pt

The complexity of human cooperation under indirect reciprocity

Fernando P. Santos[1,2,5], Jorge M. Pacheco[3,5] and Francisco C. Santos[4,5]

[1]Informatics Institute, University of Amsterdam, Science Park 904, Amsterdam 1098XH, The Netherlands
[2]Department of Ecology and Evolutionary Biology, Princeton University, Princeton, USA
[3]Centro de Biologia Molecular e Ambiental and Departamento de Matemática, Universidade do Minho, Braga 4710-057, Portugal
[4]INESC-ID and Instituto Superior Técnico, Universidade de Lisboa, IST-Taguspark, Porto Salvo 2744-016, Portugal
[5]ATP-Group, Porto Salvo P-2744-016, Portugal

FPS, 0000-0002-2310-6444; JMP, 0000-0002-2579-8499; FCS, 0000-0002-9103-2862

Indirect reciprocity (IR) is a key mechanism to understand cooperation among unrelated individuals. It involves reputations and complex information processing, arising from social interactions. By helping someone, individuals may improve their reputation, which may be shared in a population and change the predisposition of others to reciprocate in the future. The reputation of individuals depends, in turn, on social norms that define a good or bad action, offering a computational and mathematical appealing way of studying the evolution of moral systems. Over the years, theoretical and empirical research has unveiled many features of cooperation under IR, exploring norms with varying degrees of complexity and information requirements. Recent results suggest that costly reputation spread, interaction observability and empathy are determinants of cooperation under IR. Importantly, such characteristics probably impact the level of complexity and information requirements for IR to sustain cooperation. In this review, we present and discuss those recent results. We provide a synthesis of theoretical models and discuss previous conclusions through the lens of evolutionary game theory and cognitive complexity. We highlight open questions and suggest future research in this domain.

This article is part of the theme issue 'The language of cooperation: reputation and honest signalling'.

1. Introduction

From evolutionary biology to economics, cooperation in human populations has been regarded as a paradox and a challenge [1–3]. People cooperate when donating money to charity, sharing food, helping a co-worker with an arduous task or informing an outsider about the direction to a city location. Beyond small gestures, essential institutions such as social security, public health systems and courts depend on the willingness of citizens to contribute to a public good, that is, to cooperate. Despite being widespread and socially desirable, cooperation often entails individual costs to provide a benefit to others, suggesting a paradox and a challenge that can be summarized in two key questions: *How did cooperative behaviour evolve? How can cooperation be leveraged and sustained in situations where people may still defect?* Studying the evolutionary dynamics of cooperation may provide answers to these questions. In particular, learning about the evolutionary dynamics of cooperation can also unveil the reasons behind persistent defection, leaving us in a better position to solve contemporary problems by testing effective policies. Climate change [4,5], polarization [6,7], out-group hostility [8–10], corruption [11] and poverty traps [12] are certainly challenges that remind us how global cooperation (or often its absence) affects human welfare.

Several mechanisms were proposed to explain the evolution of cooperation [3]. When individuals interact repeatedly, direct reciprocity—*I cooperate with you and you cooperate with me*—is a key cooperation enabler [13–16]. Cooperation is also

THE ROYAL SOCIETY PUBLISHING

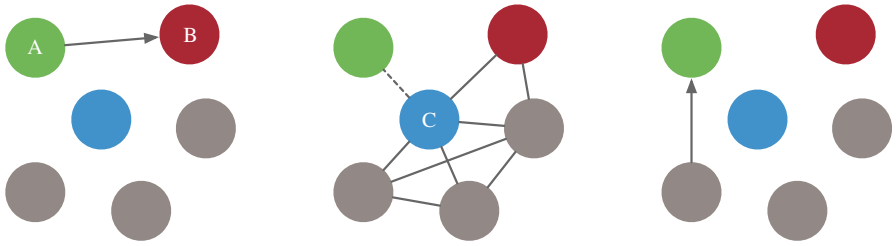

Figure 1. IR enables cooperation between unrelated individuals. (1) An individual (A, in green) cooperates with another (B, in red). (2) Within a population, observers (e.g. C, in blue) assess A's action, possibly taking into account the reputation of B, and share that information with others (e.g. gossiping to grey neighbours). (3) Information about A's previous behaviour may influence the behaviour of others with respect to A in the future. Depending on the social norms that communities employ, IR can stabilize cooperation. (Online version in colour.)

observed between unrelated individuals that did not interact directly in the past nor expect to meet again in the future. In such contexts, the human capacity to establish reputation systems provides fundamental insights into cooperation [17–25].

Reputations and social norms can work as instruments of a self-organizing process of cooperation, a combination that lies at the heart of indirect reciprocity (IR) [26–29]. Simply stated, and as summarized in figure 1, IR suggests that the action of Alice towards Bob depends on what Bob did to Carol in the past. How to judge the action of Alice? Answering that is not straightforward. Thousands of assessment rules can be mathematically formulated and used to evaluate Alice's action. Each of these rules is often called a social norm [27,30] and involves a moral judgement [26,29,31]. These norms vary in how robustly they sustain cooperation and, importantly, how complex they are [32]. The complexity of social norms, together with the information processing and communication skills required for IR, make this, perhaps, the most elaborated and cognitively demanding mechanism discovered so far. IR is said to stand in the origin of human language [27]—gossiping [33,34], a fundamental requirement for the spread of reputations, requires skilful communication mechanisms and has pervaded human social life since pre-historic times [35]. IR entails cognitive complexity (as also studied in other papers in the present theme issue [36]). But how can we quantify the complexity associated with IR? And why does complexity matter?

Here, we provide a synthesis of theoretical concepts used to understand the intricate ecologies created by reputation-based strategies and norms in IR. We provide an introduction on how reputation-based cooperation can be assessed through mathematical and computational models combining game theory and Darwin's theory of natural selection [37,38]. We navigate through the history of IR models, noting that previous works advance social norms with varying levels of cooperation and complexity. Different social norms reveal sensitivity to specific environments: costly reputation spread [39–41], interaction observability [42–44] and empathy [45,46] are examples of determinants for cooperation recently explored resorting to IR models. With a particular focus on cognitive complexity [32,36,47–49] and resorting to a complexity measure presented recently [32], we discuss those results, provide a summary of recent conclusions and point to future research avenues.

2. Modelling indirect reciprocity

The dilemma of altruistic cooperation that underlies most IR models can be captured by the so-called donation game (for exceptions see [50,51]). This game is played by pairs of individuals, one of them being the potential provider of help (donor) to the other (recipient). The donor may cooperate (C) and help the recipient at a cost c, conferring a benefit b to the recipient (with $b > c$ greater than 0). The donor can also decide not to help (defect, D), in which case no one pays any costs nor distributes any benefits. The paradox and challenges of cooperation mentioned above become clear: cooperating involves a cost ($c > 0$), yet it improves social welfare ($b > c$). For IR to solve this cooperation conundrum, information regarding individuals' actions (reputation) needs to be available in a population and donors are required to discriminate based on that information. In this regard, reputations are attributed following social norms (or assessment rules [26,52]) and reputation discrimination is implemented through cooperation strategies (or action rules [26,52])—figure 2. Each social norm defines the dynamics of reputation assessment, which in turn impacts the pay-off obtained by each strategy and, consequently, their representation in the population. Just as fit traits spread through natural selection, models of IR often resort to evolutionary game theory [37,38] and assume that strategies or norms leading to higher fitness (as measured by performance in the mentioned donation game) will spread in the population.

In theory, the complexity of social norms, that is, the amount of information processed to assess the reputation of an individual, is limitless. The simplest social norms are of (so-called) first-order, relying only on information about an individual's cooperative or defective action. Next, there are the so-called second-order norms [19,53–62]. Such norms attribute a new reputation to the donor (G or B), given information on the reputation of the recipient (G or B) and the action of the donor (C or D). This implies the existence of sixteen (up to) second-order social norms. Popular examples studied in the past are ***shunning*** [56,63] (an individual is G only if cooperates with a G opponent), ***stern-judging*** [64,65] (an individual is G if cooperates with a G opponent or defects with a B opponent) and ***simple-standing*** (SS) [54,56,66] (an individual is G if cooperates or if defects with a B opponent). Third-order social norms further require information on the reputation of the donor [30,67,68]. By including memory, it is possible to consider fourth-order social norms that also include past reputations [32]. This space of (up to) fourth-order social norms is represented in figure 2. Given these definitions of strategies and social norms, a lot of research has been devoted to understanding which norms are stable and guarantee high levels of cooperation across environments. In the last years, proposed IR systems and social

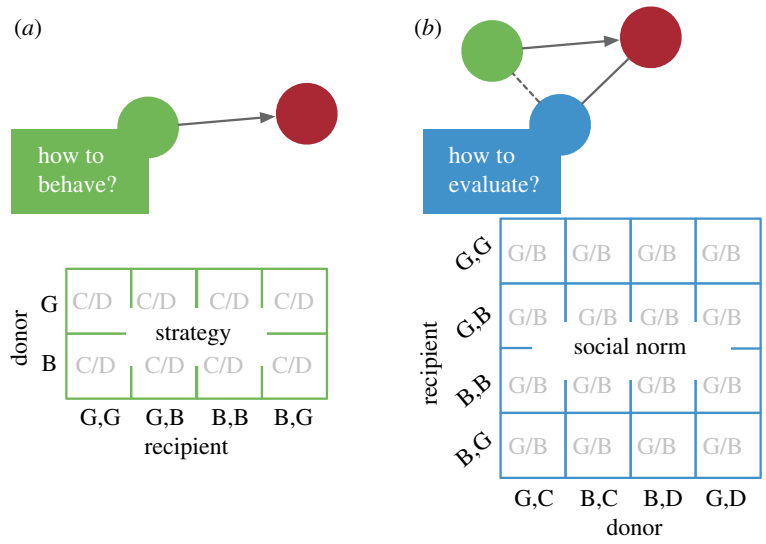

Figure 2. Norms and strategies in IR. IR involves two key decision-making rules: strategies, also known as action rules, and social norms, also known as assessment rules. As the strategy table above illustrates (*a*), in a binary world where reputations are either good (G) or bad (B), strategies discriminate based on recipients' recent and previous reputation and the donors' own reputation. Social norms (*b*) discriminate based on the donor's action (cooperate, C, or defect, D), the donor's reputation (G or B) and the recipient's recent and previous reputation (G or B). In the strategy table, the first (second) row corresponds to the action by a donor with a G (B) reputation. Each column corresponds to a pair of the current and previous reputation of the recipient. As such, the first column corresponds to a recipient that is currently G and was G in the past; the last column represents a recipient that is currently B yet was G immediately before. Likewise, in the social norm table, each entry corresponds to the new reputation of a donor (G or B) that, according to the columns, has a G or B reputation and decided to C or D with a recipient. Rows correspond to each possible pair of current and previous reputation of that recipient. Such a simplified space of strategies and social norms configures 2^8 possible strategies and 2^{16} possible social norms [32]. (Online version in colour.)

norms reveal varying degrees of complexity, as the next section elucidates.

3. A brief history of indirect reciprocity—and how complex it should be

The concept of IR was first coined by Richard Alexander [29], according to whom systems of IR synthesize the altruism of moral behaviour and the selfishness postulated by evolutionary biology. Those systems imply the continuous assessment of actions employed by members of a society, by an 'audience of interested observers'. A few years later, Boyd and Richerson employed evolutionary game theory to analyse, theoretically, the evolution of cooperation under IR [28]. They considered the existence of a chain—or stream—of interaction between individuals. An individual may help or ignore the person immediately down in that stream. In addition, individuals have a binary reputation (good or bad). IR appears here divided into two categories [52]: *vicarious reciprocity* (downstream tit-for-tat): I help someone with a good reputation, and *misguided reciprocity* (upstream tit-for-tat): I help if I previously received help. The authors concluded that downstream reciprocity is able to evolve under a wider range of circumstances than upstream reciprocity. Recently, this conclusion was validated both in online behavioural experiments [69] and through neuroimaging experiments [70]—despite evidence also supporting misguided IR in daily life situations [71]. It is important to note the different information constraints posed by these two types of IR: downstream IR requires that individuals know how others behaved in the past, whereas upstream IR simply requires information about what happened to the self. In the former case, individuals' reputation results directly from their actions: if one helps, one gets a good reputation; if one refuses to help, he/she gets a bad reputation. Only information about donors' action is required.

Following the same assessment scheme, Nowak & Sigmund [72,73] employed a set of computer simulations to study the viability of cooperation under IR. They coined the term ***image score*** (IS) to denote the reputation of an individual. As before, cooperation leads to an increase in the (public) IS, whereas defection decreases that score. The actions employed by agents are conditioned to the image level of their peers. In this set-up, the authors concluded that cooperation can be maintained, as discriminating strategies are stable to some extent. These results were validated in experiments conducted by Wedekind & Milinski [74], where individuals tended to help those who previously were generous to others. Seinen & Schram [20] confirmed that IR plays an important role in fostering cooperation, showing that participants discriminate based on information about their opponents' previous behaviours.

Under the evaluation scheme implied by IS, individuals that refuse helping peers with bad reputations will get a bad reputation. As a result, conditional cooperators will themselves be refused help, which seems at odds with evidence that humans value those that retaliate against defectors [75–78]. It was also shown that cooperation under this norm diminishes once errors [79], low mutation rates [55], low drift or high cooperation costs [54] are considered. Motivated by some of these fragilities, Leimar and Hammerstein analysed theoretically a new form of assessment, different from IS: ***standing*** [54,66]. Standing postulates that only *unjustified* defections should lead to a bad reputation. Defection is justified if the recipient has a bad reputation. Justifying defections proved to be essential to stabilize overall cooperation. Experimentally, Bolton *et al*. showed that the availability of information about opponents enables IR and, remarkably, second-order information (i.e. information about the partners' previous partners' reputation) further enhances cooperation [19]. Standing entails, however, a more

complex judgement. In this case, information about recipients' reputation is required. Subsequent laboratory experiments revealed that standing poses adoption challenges related to limited working memory and the need to keep track of a large chain of information of donors' previous partners [58].

In 2004, Ohtsuki & Iwasa [30,67] derived a comprehensive analytical apparatus to study (up to) third-order social norms that allowed for the evolutionary stability of strategies able to foster cooperation within infinite well-mixed populations. They concluded that eight norms fit that role: the so-called *leading eight* (to which group *standing* also belongs). Simultaneously, Brandt & Sigmund [79,80] studied the interplay between cooperation, defection and a discriminating strategy, under the adoption of different social norms and considering noise in the form of execution (or implementation) errors. Norms and strategies can at this point rely on three layers of information: donors' action, donors' reputation and recipient's reputation. The leading-eight norms, however, reveal a varying degree of complexity: one can find simple norms that ignore the donors' reputation (such as SS and *stern-judging*) and more complex rules (such as *standing* or ***judging***). These norms were also referred to as L1 to L8 norms [44,52]. More recently, the advantages of third-order information were stressed in [68], where authors focus on a norm named ***staying*** (that is, assess the donor following IS when the recipient is good and keep the donor reputation when the recipient is bad, regardless of the action performed). ***Consistent standing***, another leading eight, seems particularly robust in private reputation systems [44]. With consistent standing, there is only one action, in each possible context, that allows a donor to recover a good reputation; importantly, consistent standing assigns a good (bad) reputation to bad individuals that cooperate (defect) with bad (more details in figure 3). IR becomes more complex.

The idea that selection can operate on social norms suggests that different communities can employ different assessment modules, with consequent different levels of success in sustaining cooperation. Pacheco *et al*. analysed a multi-level model where norms and strategies co-evolve at different scales [64,83]: intra-tribe strategy dynamics is combined with an inter-tribe social norm proliferation. While the successful strategies are imitated within a tribe, the social norms that confer high levels of fitness allow one tribe to impose this norm over others. Importantly, this model (as well as others [44,84–86]) allows us to study not only cooperation levels under a fixed social norm, but also how different norms—and the own IR system—can evolve. It was shown that ***stern-judging***, one of the leading-eight norms, is able to emerge as a successful norm. While allowing for the emergence of (up to) third-order assessment rules, the prevailing one is merely of second-order: complex norms do not necessarily translate into success [32].

4. Recent and future research

The previous results reveal an important and significant research effort dedicated to uncovering the social norms that can evolve and sustain cooperation across different domains. Social norms with increasing complexity are discussed. Importantly, recent results stress how particular settings affect the efficiency of norms. Different environment and individuals' characteristics probably impact the level of complexity and information required for IR to sustain cooperative behaviours. Of particular importance are costly reputation spread, interaction observability and empathy. Each of these characteristics suggests new research avenues, where, as we shall discuss, it becomes relevant to consider the complexity and information requirements associated with IR.

(a) Costly reputation spread

Despite constituting a promising way of eliciting cooperation, IR requires that individuals share their experiences with others, a process that may be costly. Most previous models assume that reputation spread depends on exogenous factors [87]. In reality, however, accessing the information about a private interaction relies on the decision of the individuals involved who may be willing to share (or not) the information. For example, in *e-commerce* or on *online community* platforms [88], private interactions occur, and individuals are supposed to provide information about their opponents' actions. When information sharing is costly (cost here may simply translate in time spent providing such information), and devoid of any incentives, reporting is hardly fulfilled by rational agents, such that the system of IR—and consequently cooperation—may collapse. Moreover, situations where individuals may profit from deceitful communication may further undermine cooperation under IR, as discussed by Számadó *et al*. [62]; in these contexts it is fundamental to understand how to maintain honest gossiping (as investigated in the present theme issue [89–92]). Reputation spread constitutes a second-order free-rider problem: everyone would benefit from IR, yet its maintenance is costly, and it is tempting to avoid the burden [3]. This dilemma is highlighted in [40]. In this context, Sasaki *et al.* show that pre-assessment of individuals that fail to share reputations can stabilize both cooperation and the willingness to contribute to the costly reputation system [37]. Likewise, in [39], it is shown that anticipating which individuals in the population are willing to pay a cost to share reputations provides an escape to the second-order dilemma of costly reputation spread. Once again, assessing complexity in IR is likely to play an important role in this domain: (i) how do complex social norms (and strategies) impact the willingness of an individual to follow them to share the reputations of others? and (ii) how to guarantee that individuals can anticipate that others are likely to share the outcome of their interactions?

(b) Interaction observability and private interactions

Private reputation systems, and the possibility that individuals disagree in what concerns others' reputations are relevant aspects that can affect cooperation under particular norms. This was noted by Uchida & Sasaki [42,43], showing that, particularly under stern-judging, errors in reputation assessment can lead to disagreements in the population about individuals' reputations, which can ultimately undermine cooperation. On top of disagreements, private reputations can also underlie the emergence of antagonist groups [93]. Such disagreements in moral judgements can result from differences in cultures and contexts, as discussed in the present theme issue [94]. The challenges associated with private reputations were also recently studied by Hilbe *et al*., in a work revealing that most leading-eight social norms lose stability in the context of private reputations systems [44]. *Consistent standing*, one of the most complex leading-eight norms (figure 3) is the most stable norm in this context [44]. This research line also stresses the need to explicitly consider complexity measures in IR: (i) given that reputation

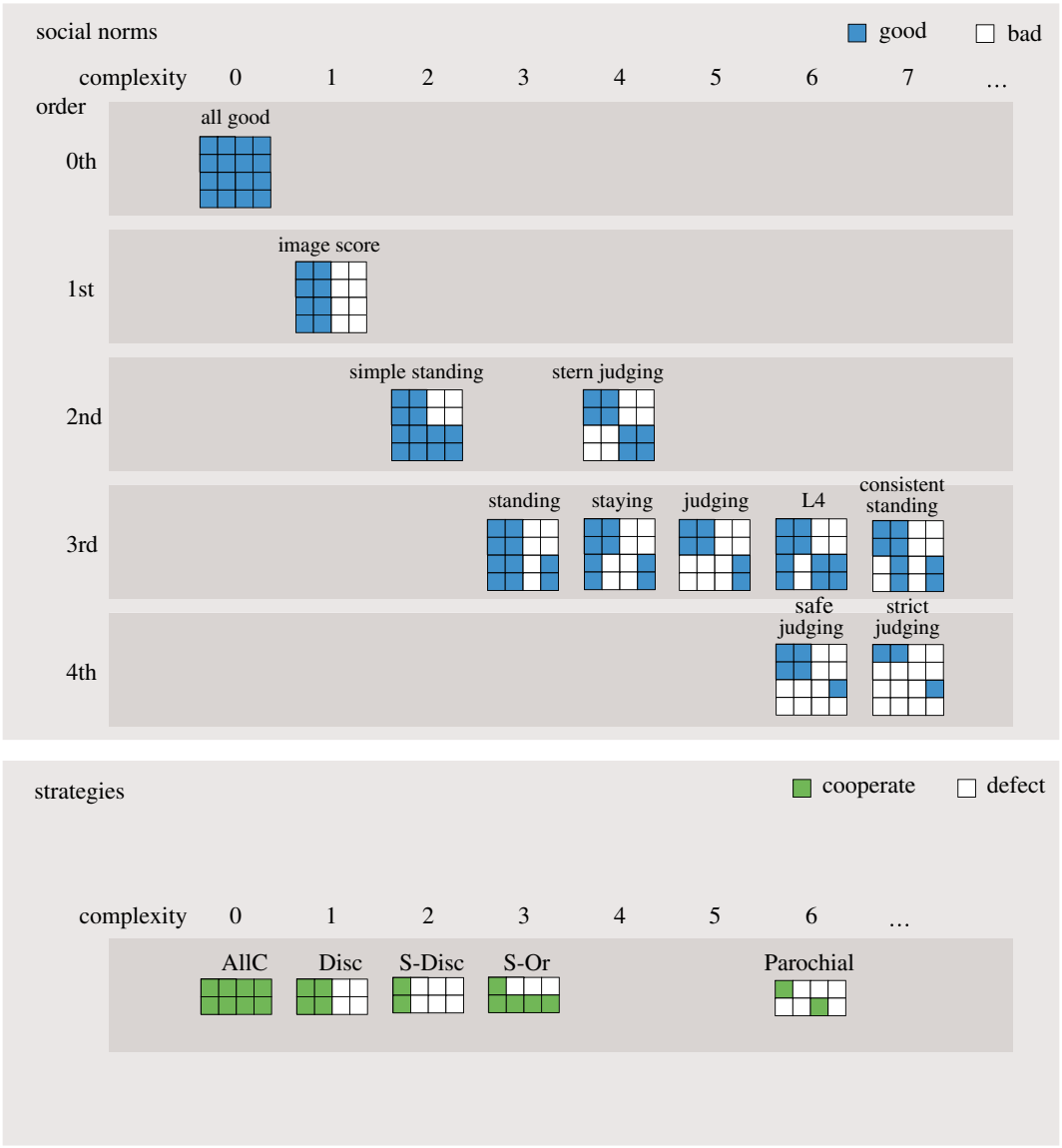

Figure 3. The complexity of strategies and social norms. Under IR, and even in a world of binary reputations and decisions, social norms and strategies can rely on arbitrarily large information sets and encapsulate arbitrarily complex decision-making rules. The level of information used can be captured by the *order* of social norms, which translates the number of information layers that a norm can use: 0th-order norms prescribe unconditional judgements, such as everyone is good (all good); 1st-order norms discriminate based on the donor's action—such as the well-known IS norm that assigns G only when donors cooperate; 2nd-order norms include information about the recipient's reputation; 3rd-order norms include information about the donor's reputation; and finally, 4th-order norms include information about the previous reputation of the recipient [32]. Naturally, other sources of information could be considered (e.g. [81]), possibly expanding a norms' order beyond the 4th. Order is represented by the rows in the panel above. Within each set of norms with a given order, information can be combined to formulate judgements with a variable degree of complexity (columns). Relying on the binary value of reputations and actions, we capture complexity through the so-called *Boolean complexity* [32,48]. In the figure, we provide a visualization according to the layout of strategy and social norm tables represented in figure 2. Following the method of *Karnaugh maps* [82], the complexity of norms and strategies can be determined by counting the number of blocks of size 2^k 'G's or 'Cs', until all coloured cells are covered and where k is chosen to be as large as possible and blocks can overlap: each block of 2^k 'G's or 'Cs' increases the complexity of norms by $4-k$ and strategies by $3-k$. (Online version in colour.)

disagreement is harmful, how likely are they to exist as a function of social norm complexity? (ii) private reputation systems assume that each individual can possibly keep a different record on everyone else in the population—how is that impacted by limited memory and cognitive abilities? and (iii) what type of synchronization mechanisms (e.g. gossip [33,34,89–92] or broadcast institutions [95]) can be both simple to be widely used and contribute for reputations to become homogeneous?

(c) Empathy

The previous challenges of private reputation systems are alleviated if individuals are empathetic when judging others, as recently pointed out by Radzvilavicius *et al*. [45]. Empathy, here, means that individuals are able to place themselves in the donors' shoes, when judging their behaviours. Technically, this implies that observers use the information that donors had when they decided to cooperate or defect with the recipient. In opposition, egocentric observers use their own information to judge donors. Judging donors empathetically contributes to reducing the disagreements that undermine cooperation in private reputation systems. Considering individuals' cognitive abilities explicitly and measuring IR complexity seems here fruitful, as well: how to inspire empathetic judgements, knowing that requires that individuals place themselves in others' shoes and

consider their perspectives—which calls for high cognitive abilities, such as theory of mind?

Over the last 30 years, works on IR debate social norms that reveal particular strengths and limitations in different environments and, importantly, are characterized by varying degrees of information requirements and complexity. Measuring IR's complexity requirements is likely to inspire new research at the forefront of IR modelling, particularly in the domains of costly reputation spread, private reputation systems and empathetic judgements. Furthermore, the performance of a complex social norm may be constrained by human's difficulty to follow complex subjective rules [47–49,96]. A fundamental aspect should then be discussed: how to measure complexity in IR?

5. Complexity of human cooperation under indirect reciprocity

Measuring complexity in IR poses empirical and technical challenges. On the one hand, it requires understanding how humans learn and apply subjective logic rules and which rule-representations can capture the characteristics inherent to that process. On the other hand, it is desirable to adopt a measure of complexity that can systematically be applied to the large space of social norms and strategies in IR (as in fourth order, figure 2). The order of a norm (as introduced in §2) can be conceived as an incipient complexity measure: first-order norms require less (and/or more readily accessible) information than, say, second- or third-order norms. Within a given order, however, norms can vary in how complexly information is used to assign new reputations.

In economics, in the context of repeated games, the complexity of strategies is often captured by translating them to state automata and counting the number of elements (states or transitions) in such representation [47,97–100]. In fact, recent results reveal that state-complexity (i.e. number of states in the corresponding automata) translate how difficult it is for humans to use a given strategy [47]. The difficulty of using a complex strategy is typically operationalized through *complexity costs*, associating strategy usage with a cost proportional to its complexity [47,98,100]. Technically, however, it can become unfeasible to systematically translate all IR social norms to an automaton and visually count its states.

An alternative consists in regarding norms and strategies as Boolean functions that, when evaluated as *true*, will assign a good reputation or cooperate [32]. This approach naturally takes advantage of the fact that models of IR often consider binary reputations and binary actions. As a result, each layer of information used by strategies and norms can be conceived as a Boolean (logic) variable that can acquire two values: good/cooperate (*true*) or bad/defect (*false*). In the following, we assume that the (A)ction of the donor is represented by variable A, the current (R)eputation of the (D)onor is represented by R_D, the (A)ctual (R)eputation of the recipient is represented by R_A and the (P)revious (R)eputation of the recipient is captured by R_P. Again, these variables have value *true* (good or cooperate) or *false* (bad or defect). Let us exemplify this formalism: IS, a norm postulating that those that cooperate are good and those that defect are bad, can be represented through the Boolean function $f_{IS}(A, R_D, R_A, R_P) = A$. Note that if A is true (that is, the action of the donor is C), the norm is evaluated to true, which means that a G reputation is attributed. Likewise, the norm SS, postulating that an individual is good if she cooperates or if she defects against those that are bad, can be written as $f_{SS}(A, R_D, R_A, R_P) = A \vee \bar{A} \wedge \overline{R_A}$, where $\bar{A}$ represents the negation of A, $\vee$ represents the logic disjunction (*or*) and $\wedge$ the logic conjunction (*and*). Following the same principle, this function assigns a good reputation if the donor cooperates (A) or if the donor defects $\bar{A}$ and the recipient is currently bad ($\overline{R_A}$). Similarly, strategies can also be written as logic formulae. The strategy Disc (discriminator, that is, cooperate with the good and defect with the bad) can be written as $f_{\mathrm{Disc}}(R_D, R_A, R_P) = R_A$. A stricter strategy, named strict-discriminator (S-Disc), that only cooperates with a recipient that is good now and was good in the past, can be written as $f_{S-Disc}(R_D, R_A, R_P) = R_A \wedge R_P$.

Using this formalism, the complexity of norms and strategies can conveniently be computed as the number of literals (that is, logic variables or their negation) in their minimal logic formula. This way, $f_{SS}(A, R_D, R_A, R_P) = A \vee \bar{A} \wedge \overline{R_A}$, has three literals ($A$, $\bar{A}$ and $\overline{R_A}$) and thereby a complexity of 3. $f_{S-Disc}(R_D, R_A, R_P) = R_A \wedge R_P$ has two literals (R_A and R_P) and a complexity of 2. Importantly, this way of quantifying complexity, also known as *Boolean complexity*, was shown to provide a relatively good heuristic of how easily humans can learn Boolean concepts [48] (with some limitations pointed in [101]). Here, we equate complexity to the number of literals in the minimal disjunctive normal form of a norm/strategy (see [32] or [101] for a step-by-step guide of how to minimize logic formulae). Though not unique, this choice is technically convenient: we can apply readily available methods (e.g. the *Quine–McCluskey* algorithm [101–103] or *Karnaugh maps* [82]) and software implementations.

Figure 3 explicitly organizes strategies and norms in terms of their order (rows) and complexity (columns). We represent well-known third-order norms, most of them already alluded to in §2—*standing* and *judging* [27,67], *staying* [68], L4 [44,52] and *consistent standing* [44]—as well as the most cooperative second-order norms—SS and *stern-judging* [53,55,57]. We also highlight two fourth-order extensions of *judging*: *safe judging*, that only assigns G to B donors if they defect with recipients that are consistently B (by consistent we mean here that present and past reputations are the same); and *strict judging*, that only assigns G to donors that cooperate with recipients that are consistently G. Social norms are organized in a tabular format (figure 2), which provides a compact representation while offering a means to grasp, visually, their order and complexity. Norms of first order have all entries of the left and the right eight-entry blocks identical. Norms of second-order have four four-entry blocks in each of which all entries are equal; norms of third-order have eight two-entry blocks of this type.

Strategies are represented following a similar principle. In figure 3, we represent the trivial, unconditional, strategy *always cooperate* (AllC), the strategy *discriminator* (Disc, cooperate with G and defect with B), *strict-discriminator* (S-Disc, cooperate only with those consistently G), *strict-or* (S-Or, cooperate if have a reputation B or if the opponent was consistently G) and Parochial (cooperate if my opponent consistently has the exact same reputation as me).

Measuring the complexity of social norms and strategies allows us to apprehend which norms sustain high cooperation levels while keeping simplicity. Other advantages of capturing the complexity of strategies under IR include: (i) understanding how social norms vary in the

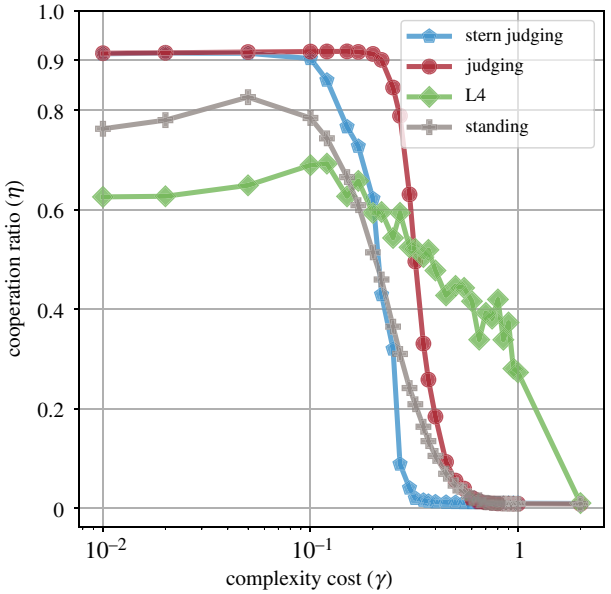

Figure 4. Strategies' complexity costs and cooperation under different social norms. Here, we represent the cooperation ratio (fraction of interactions that lead to cooperation) given different social norms and complexity costs associated with each strategy. We assume that individuals pay a cost proportional to the complexity of each strategy (γ), on top of the costs/benefits associated with the donation game played. We compare cooperation under *judging* (red, circles), *stern-judging* (blue, pentagons), *standing* (grey, plus symbol) and L4 (green, diamonds). We consider a population of $Z = 50$ individuals that can adopt any of the strategies represented in figure 2. Initially, strategies are attributed at random and, at each time-step, pairs of individuals (say, A and B) are sampled; A adopts B's strategy with probability $1/(1 + e^{f_A - f_B})$, where f_X is the average pay-off of individual X obtained after playing against Z random opponents [106]. After each interaction, public reputations are updated following the corresponding social norm. With probability $\mu = 1/Z$ individuals adopt a random strategy (mutation). Each run lasts for $10^5 Z$ time-steps and results correspond to averages over the last 80% of time-steps of 200 runs. Execution, assignment and assessment errors occur with a probability 0.01 [53]. Other parameters: $b = 5$, $c = 1$. (Online version in colour.)

complexity of behaviours that they sustain, and (ii) how robust is cooperation under a given norm when individuals reveal difficulties in implementing complex strategies. The feasibility of employing strategies of varying complexity in repeated interactions has been operationalized through the introduction of complexity costs [32,47,79,98,104,105]. Analogously, we provide in figure 4 a simple proof of concept of how cooperation under IR depends on the feasibility of individuals to implement complex strategies.

To this end, we consider that individuals play a donation game, possibly adopting any of the strategies in the strategy-space represented in figure 2. Besides the pay-off resulting from the donation game, individuals using a strategy p will pay a cost $\gamma\kappa(p)$, where $\kappa(p)$ is the complexity associated with strategy p and γ is the complexity cost factor considered. Figure 4 represents the cooperation ratio (i.e. fraction of actions that result in cooperation) for four prototypical social norms, defined in figure 3. We observe that norms react differently to complexity costs: some norms lead to high levels of cooperation in the limit of no complexity costs ($\gamma = 0$) (*standing*, *judging* and *stern-judging*), whereas other norms promote intermediate levels of cooperation even when individuals may find harder to implement more complex strategies (i.e. higher values of γ). The effectiveness of social norms in promoting cooperation depends on individuals' capacity to implement complex strategies.

6. Discussion

In this paper, we review models and results on cooperation under IR. We highlight works that, over the last 30 years, debated which social norms—and in which contexts—sustain high levels of cooperation. Multiple excellent reviews provide summaries of research in IR [26,27,107,108]. Here, we stress the levels of complexity and information requirements implied in previous IR works. By adopting a definition of social norm and strategy complexity based on Boolean complexity [32], we discuss previously proposed norms in terms of their order, complexity and cooperation level. Finally, we provide a proof of concept that emphasizes how complexity costs in IR can fundamentally alter the ranking of most cooperative social norms: different assessment rules perform differently, depending on individuals' ability to employ complex strategies.

We argue that measuring complexity in IR is advantageous for three extra reasons. First, it can guide research towards a better understanding of which social norms and strategies (that sustain cooperation) are likely to be (i) used by humans at large, (ii) communicated easily and (iii) applied without mistakes and prevail. After all, simple principles—such as The Golden Rule—seem to stand the test of time. Second, in line with *Occam's razor* principle, measuring complexity can contribute to identify the simplest explanations for the evolution of cooperation under IR. Third, in the context of online reputation systems—an area often associated with IR [27,88,109]—simple norms can inspire easier rules to attribute reputations, which are also easier to explain to users and, potentially, to improve compliance.

Details on the complexity of social norms and strategies under IR are provided in [32]. Departing from the concepts here discussed, different directions can be followed in order to improve our understanding of human complexity in IR. First of all, the theoretical results presented suggest that new field and laboratory experiments are carried out to better understand the role of complexity costs in human decision-making and reciprocity (such as in [47]). Second, new experiments would also be beneficial to apprehend how variations in strategy/norm logic representation, and corresponding complexity measures, can better capture the human difficulties in implementing specific behaviours in IR. On top of the mentioned literature on finite state automata [47,97,98,100], other alternatives include extended logic formalisms [49], alternative logic operators [110] or other measures of Boolean function complexity [111]. Third, it can be relevant to extend the formalism we discuss in the present paper to quantify cognitive complexity in other domains (beyond cooperation) where social interactions are also mediated by reputations [112].

Data accessibility. Code to calculate Boolean complexity is provided in https://doi.org/10.5281/zenodo.1041379. We follow the same simulation procedure described in https://doi.org/10.1038/nature25763 [32].

Authors' contributions. F.P.S., J.M.P. and F.C.S. conceived the study, designed the study, wrote the manuscript and gave final approval for publication.

Competing interests. We declare we have no competing interests.

Funding. The authors acknowledge the support from FCT-Portugal (grant nos. PTDC/MAT-APL/6804/2020, UIDB/04050/2020, UIDB/50021/2020 and PTDC/CCI-INF/7366/2020).

Acknowledgements. F.P.S. acknowledges support from the James S. McDonnell Foundation twenty-first Century Science Initiative in Understanding Dynamic and Multi-scale Systems Postdoctoral Fellowship Award.

References

1. Pennisi E. 2009 On the origin of cooperation. *Science* **325**, 1196–1199. (doi:10.1126/science.325_1196)
2. Fehr E, Fischbacher U. 2003 The nature of human altruism. *Nature* **425**, 785–791. (doi:10.1038/nature02043)
3. Rand DG, Nowak MA. 2013 Human cooperation. *Trends Cogn. Sci.* **17**, 413–425.
4. Vasconcelos VV, Santos FC, Pacheco JM, Levin SA. 2014 Climate policies under wealth inequality. *Proc. Natl Acad. Sci. USA* **111**, 2212–2216. (doi:10.1073/pnas.1323479111)
5. Tavoni A, Levin S. 2014 Managing the climate commons at the nexus of ecology, behaviour and economics. *Nat. Clim. Change* **4**, 1057–1063. (doi:10.1038/nclimate2375)
6. Iyengar S, Lelkes Y, Levendusky M, Malhotra N, Westwood SJ. 2019 The origins and consequences of affective polarization in the United States. *Ann. Rev. Political Sci.* **22**, 129–146. (doi:10.1146/annurev-polisci-051117-073034)
7. Finkel EJ *et al.* 2020 Political sectarianism in America. *Science* **370**, 533–536.
8. Choi J-K, Bowles S. 2007 The coevolution of parochial altruism and war. *Science* **318**, 636–640. (doi:10.1126/science.1144237)
9. Gross J, De Dreu CK. 2019 The rise and fall of cooperation through reputation and group polarization. *Nat. Commun.* **10**, 1–10. (doi:10.1038/s41467-019-08727-8)
10. Whitaker RM, Colombo GB, Rand DG. 2018 Indirect reciprocity and the evolution of prejudicial groups. *Sci. Rep.* **8**, 1–14. (doi:10.1038/s41598-018-31363-z)
11. Lee J-H, Iwasa Y, Dieckmann U, Sigmund K. 2019 Social evolution leads to persistent corruption. *Proc. Natl Acad. Sci. USA* **116**, 13 276–13 281. (doi:10.1073/pnas.1900078116)
12. Santos FP, Pacheco JM, Santos FC, Levin SA. 2021 Dynamics of informal risk sharing in collective index insurance. *Nat. Sustain.* **4**, 1–7. (doi:10.1038/s41893-020-00667-2)
13. Trivers RL. 1971 The evolution of reciprocal altruism. *Q. Rev. Biol.* **46**, 35–57. (doi:10.1086/406755)
14. Axelrod R, Hamilton WD. 1981 The evolution of cooperation. *Science* **211**, 1390–1396. (doi:10.1126/science.7466396)
15. Hilbe C, Chatterjee K, Nowak MA. 2018 Partners and rivals in direct reciprocity. *Nat. Hum. Behav.* **2**, 469–477. (doi:10.1038/s41562-018-0320-9)
16. Power EA, Ready E. 2018 Building bigness: reputation, prominence, and social capital in rural South India. *Am. Anthropol.* **120**, 444–459. (doi:10.1111/aman.13100)
17. Wu J, Balliet D, Van Lange PA. 2016 Reputation, gossip, and human cooperation. *Soc. Pers. Psychol. Compass* **10**, 350–364. (doi:10.1111/spc3.12255)
18. Milinski M, Semmann D, Krambeck H-J. 2002 Reputation helps solve the 'tragedy of the commons'. *Nature* **415**, 424–426. (doi:10.1038/415424a)
19. Bolton GE, Katok E, Ockenfels A. 2005 Cooperation among strangers with limited information about reputation. *J. Public Econ.* **89**, 1457–1468. (doi:10.1016/j.jpubeco.2004.03.008)
20. Seinen I, Schram A. 2006 Social status and group norms: indirect reciprocity in a repeated helping experiment. *Euro. Econ. Rev.* **50**, 581–602. (doi:10.1016/j.euroecorev.2004.10.005)
21. Pfeiffer T, Tran L, Krumme C, Rand DG. 2012 The value of reputation. *J. R. Soc. Interface* **9**, 2791–2797. (doi:10.1098/rsif.2012.0332)
22. Giardini F, Wittek R. 2019 Gossip, reputation, and sustainable cooperation: sociological foundations. In *The Oxford handbook of gossip and reputation* (eds F Giardini, R Wittek), pp. 23–46. Oxford, UK: Oxford University Press.
23. Samu F, Számadó S, Takács K. 2020 Scarce and directly beneficial reputations support cooperation. *Sci. Rep.* **10**, 1–12. (doi:10.1038/s41598-020-68123-x)
24. Giardini F, Conte R. 2012 Gossip for social control in natural and artificial societies. *Simulation* **88**, 18–32. (doi:10.1177/0037549711406912)
25. Roberts G, Raihani N, Bshary R, Manrique HM, Farina A, Samu F, Barclay P. 2021 The benefits of being seen to help others: indirect reciprocity and reputation-based partner choice. *Phil. Trans. R. Soc. B* **376**, 20200290. (doi:10.1098/rstb.2020.0290)
26. Sigmund K. 2012 Moral assessment in indirect reciprocity. *J. Theor. Biol.* **299**, 25–30. (doi:10.1016/j.jtbi.2011.03.024)
27. Nowak MA, Sigmund K. 2005 Evolution of indirect reciprocity. *Nature* **437**, 1291–1298. (doi:10.1038/nature04131)
28. Boyd R, Richerson PJ. 1989 The evolution of indirect reciprocity. *Soc. Net.* **11**, 213–236. (doi:10.1016/0378-8733(89)90003-8)
29. Alexander RD. 1987 *The biology of moral systems*. New York, NY: Transaction Publishers.
30. Ohtsuki H, Iwasa Y. 2004 How should we define goodness?—reputation dynamics in indirect reciprocity. *J. Theor. Biol.* **231**, 107–120. (doi:10.1016/j.jtbi.2004.06.005)
31. Efferson C, Fehr E. 2018 Simple moral code supports cooperation. *Nature* **555**, 169–170. (doi:10.1038/d41586-018-02621-x)
32. Santos FP, Santos FC, Pacheco JM. 2018 Social norm complexity and past reputations in the evolution of cooperation. *Nature* **555**, 242–245. (doi:10.1038/nature25763)
33. Dores Cruz TD *et al.* 2021 Gossip and reputation in everyday life. *Phil. Trans. R. Soc. B* **376**, 20200301. (doi:10.1098/rstb.2020.0301)
34. Hess NH, Hagen EH. 2021 Competitive gossip: the impact of domain, resource value, resource scarcity and coalitions. *Phil. Trans. R. Soc. B* **376**, 20200305. (doi:10.1098/rstb.2020.0305)
35. Dunbar R, Dunbar RIM. 1998 *Grooming, gossip, and the evolution of language*. Cambridge, MA: Harvard University Press.
36. Manrique HM, Zeidler H, Roberts G, Barclay P, Walker M, Samu F, Fariña A, Bshary R, Raihani N. 2021 The psychological foundations of reputation-based cooperation. *Phil. Trans. R. Soc. B* **376**, 20200287. (doi:10.1098/rstb.2020.0287)
37. Weibull JW. 1997 *Evolutionary game theory*. Cambridge, MA: MIT press.
38. Nowak MA. 2006 *Evolutionary dynamics: exploring the equations of life*. Cambridge, MA: Harvard University Press.
39. Santos F, Pacheco J, Santos F. 2018 Social norms of cooperation with costly reputation building. In *Proc. of the AAAI Conf. on Artificial Intelligence*, Vol. 32. Palo Alto, CA: AAAI Press.
40. Suzuki S, Kimura H. 2013 Indirect reciprocity is sensitive to costs of information transfer. *Sci. Rep.* **3**, 1–5.
41. Sasaki T, Okada I, Nakai Y. 2016 Indirect reciprocity can overcome free-rider problems on costly moral assessment. *Biol. Lett.* **12**, 20160341. (doi:10.1098/rsbl.2016.0341)
42. Uchida S. 2010 Effect of private information on indirect reciprocity. *Phys. Rev. E* **82**, 036111. (doi:10.1103/PhysRevE.82.036111)
43. Uchida S, Sasaki T. 2013 Effect of assessment error and private information on stern-judging in indirect reciprocity. *Chaos Solitons Fractals* **56**, 175–180. (doi:10.1016/j.chaos.2013.08.006)
44. Hilbe C, Schmid L, Tkadlec J, Chatterjee K, Nowak MA. 2018 Indirect reciprocity with private, noisy, and incomplete information. *Proc. Natl Acad. Sci. USA* **115**, 12 241–12 246. (doi:10.1073/pnas.1810565115)
45. Radzvilavicius AL, Stewart AJ, Plotkin JB. 2019 Evolution of empathetic moral evaluation. *eLife* **8**, e44269. (doi:10.7554/eLife.44269)
46. Masuda N, Santos FC. 2019 A mathematical look at empathy. *eLife* **8**, e47036. (doi:10.7554/eLife.47036)
47. Oprea R. 2020 What makes a rule complex? *Am. Econ. Rev.* **110**, 3913–3951. (doi:10.1257/aer.20191717)
48. Feldman J. 2000 Minimization of Boolean complexity in human concept learning. *Nature* **407**, 630–633. (doi:10.1038/35036586)
49. Jamroga W, Malvone V, Murano A. 2019 Natural strategic ability. *Artificial Intelligence* **277**, 103170. (doi:10.1016/j.artint.2019.103170)
50. Clark D, Fudenberg D, Wolitzky A. 2020 Indirect reciprocity with simple records. *Proc. Natl Acad. Sci. USA* **117**, 11 344–11 349. (doi:10.1073/pnas.1921984117)
51. Nakamura M, Ohtsuki H. 2014 Indirect reciprocity in three types of social dilemmas. *J. Theor. Biol.* **355**, 117–127. (doi:10.1016/j.jtbi.2014.03.035)
52. Sigmund K. 2016 *The calculus of selfishness*. Princeton, NJ: Princeton University Press.
53. Santos FP, Santos FC, Pacheco JM. 2016 Social norms of cooperation in small-scale societies. *PLoS Comput. Biol.* **12**, e1004709. (doi:10.1371/journal.pcbi.1004709)
54. Leimar O, Hammerstein P. 2001 Evolution of cooperation through indirect reciprocity. *Proc. R. Soc. B* **268**, 745–753. (doi:10.1098/rspb.2000.1573)
55. Santos FP, Pacheco JM, Santos FC. 2016 Evolution of cooperation under indirect reciprocity and arbitrary

exploration rates. *Sci. Rep.* **6**, 1–9. (doi:10.1038/s41598-016-0001-8)

56. Panchanathan K, Boyd R. 2003 A tale of two defectors: the importance of standing for evolution of indirect reciprocity. *J. Theor. Biol.* **224**, 115–126. (doi:10.1016/S0022-5193(03)00154-1)
57. Ohtsuki H, Iwasa Y. 2007 Global analyses of evolutionary dynamics and exhaustive search for social norms that maintain cooperation by reputation. *J. Theor. Biol.* **244**, 518–531. (doi:10.1016/j.jtbi.2006.08.018)
58. Milinski M, Semmann D, Bakker TC, Krambeck H-J. 2001 Cooperation through indirect reciprocity: image scoring or standing strategy? *Proc. R. Soc. B* **268**, 2495–2501. (doi:10.1098/rspb.2001.1809)
59. Engelmann D, Fischbacher U. 2009 Indirect reciprocity and strategic reputation building in an experimental helping game. *Games Econ. Behav.* **67**, 399–407. (doi:10.1016/j.geb.2008.12.006)
60. Swakman V, Molleman L, Ule A, Egas M. 2016 Reputation-based cooperation: empirical evidence for behavioral strategies. *Evol. Hum. Behav.* **37**, 230–235. (doi:10.1016/j.evolhumbehav.2015.12.001)
61. Ohtsuki H, Iwasa Y, Nowak MA. 2009 Indirect reciprocity provides only a narrow margin of efficiency for costly punishment. *Nature* **457**, 79–82. (doi:10.1038/nature07601)
62. Számadó S, Szalai F, Scheuring I. 2016 Deception undermines the stability of cooperation in games of indirect reciprocity. *PLoS ONE* **11**, e0147623. (doi:10.1371/journal.pone.0147623)
63. Takahashi N, Mashima R. 2006 The importance of subjectivity in perceptual errors on the emergence of indirect reciprocity. *J. Theor. Biol.* **243**, 418–436. (doi:10.1016/j.jtbi.2006.05.014)
64. Pacheco JM, Santos FC, Chalub FAC. 2006 Stern-judging: a simple, successful norm which promotes cooperation under indirect reciprocity. *PLoS Comput. Biol.* **2**, e178. (doi:10.1371/journal.pcbi.0020178)
65. Kandori M. 1992 Social norms and community enforcement. *Rev. Econ. Stud.* **59**, 63–80. (doi:10.2307/2297925)
66. Sugden R. 2004 *The economics of rights, co-operation and welfare*. Berlin, Germany: Springer.
67. Ohtsuki H, Iwasa Y. 2006 The leading eight: social norms that can maintain cooperation by indirect reciprocity. *J. Theor. Biol.* **239**, 435–444. (doi:10.1016/j.jtbi.2005.08.008)
68. Sasaki T, Okada I, Nakai Y. 2017 The evolution of conditional moral assessment in indirect reciprocity. *Sci. Rep.* **7**, 1–8. (doi:10.1038/srep41870)
69. Capraro V, Marcelletti A. 2014 Do good actions inspire good actions in others? *Sci. Rep.* **4**, 1–6. (doi:10.1038/srep07470)
70. Watanabe T, Takezawa M, Nakawake Y, Kunimatsu A, Yamasue H, Nakamura M, Miyashita Y, Masuda N. 2014 Two distinct neural mechanisms underlying indirect reciprocity. *Proc. Natl Acad. Sci. USA* **111**, 3990–3995. (doi:10.1073/pnas.1318570111)
71. Mujcic R, Leibbrandt A. 2018 Indirect reciprocity and prosocial behaviour: evidence from a natural field experiment. *Econ. J.* **128**, 1683–1699. (doi:10.1111/ecoj.12474)
72. Nowak MA, Sigmund K. 1998 Evolution of indirect reciprocity by image scoring. *Nature* **393**, 573–577. (doi:10.1038/31225)
73. Nowak MA, Sigmund K. 1998 The dynamics of indirect reciprocity. *J. Theor. Biol.* **194**, 561–574. (doi:10.1006/jtbi.1998.0775)
74. Wedekind C, Milinski M. 2000 Cooperation through image scoring in humans. *Science* **288**, 850–852. (doi:10.1126/science.288.5467.850)
75. Boyd R, Richerson PJ. 1992 Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethol. Sociobiol.* **13**, 171–195. (doi:10.1016/0162-3095(92)90032-Y)
76. dos Santos M, Rankin DJ, Wedekind C. 2013 Human cooperation based on punishment reputation. *Evolution* **67**, 2446–2450. (doi:10.1111/evo.12108)
77. Fehr E, Gächter S. 2002 Altruistic punishment in humans. *Nature* **415**, 137–140. (doi:10.1038/415137a)
78. dos Santos M, Wedekind C. 2015 Reputation based on punishment rather than generosity allows for evolution of cooperation in sizable groups. *Evol. Hum. Behav.* **36**, 59–64. (doi:10.1016/j.evolhumbehav.2014.09.001)
79. Brandt H, Sigmund K. 2006 The good, the bad and the discriminator—errors in direct and indirect reciprocity. *J. Theor. Biol.* **239**, 183–194. (doi:10.1016/j.jtbi.2005.08.045)
80. Brandt H, Sigmund K. 2005 Indirect reciprocity, image scoring, and moral hazard. *Proc. Natl Acad. Sci. USA* **102**, 2666–2670. (doi:10.1073/pnas.0407370102)
81. de Melo CM, Terada K, Santos FC. 2021 Emotion expressions shape human social norms and reputations. *iScience* **24**, 102141. (doi:10.1016/j.isci.2021.102141)
82. Karnaugh M. 1953 The map method for synthesis of combinational logic circuits. *Trans. Am. Inst. Electrical Eng. Part I* **72**, 593–599. (doi:10.1109/TCE.1953.6371932)
83. Chalub FA, Santos FC, Pacheco JM. 2006 The evolution of norms. *J. Theor. Biol.* **241**, 233–240. (doi:10.1016/j.jtbi.2005.11.028)
84. Brandt H, Sigmund K. 2004 The logic of reprobation: assessment and action rules for indirect reciprocation. *J. Theor. Biol.* **231**, 475–486. (doi:10.1016/j.jtbi.2004.06.032)
85. Xu J, García J, Handfield T. 2019 Cooperation with bottom-up reputation dynamics. In *Proc. of the 18th Int. Conf. on Autonomous Agents and MultiAgent Systems*, pp. 269–276. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.
86. Yamamoto H, Okada I, Uchida S, Sasaki T. 2017 A norm knockout method on indirect reciprocity to reveal indispensable norms. *Sci. Rep.* **7**, 1–7. (doi:10.1038/srep44146)
87. Ohtsuki H, Iwasa Y, Nowak MA. 2015 Reputation effects in public and private interactions. *PLoS Comput. Biol.* **11**, e1004527. (doi:10.1371/journal.pcbi.1004527)
88. Resnick P, Zeckhauser R, Swanson J, Lockwood K. 2006 The value of reputation on eBay: a controlled experiment. *Exp. Econ.* **9**, 79–101. (doi:10.1007/s10683-006-4309-2)
89. Fonseca MA, Peters K. 2021 Is it costly to deceive? People are adept at detecting gossipers' lies but may not reward honesty. *Phil. Trans. R. Soc. B* **376**, 20200304. (doi:10.1098/rstb.2020.0304)
90. Giardini F, Vilone D, Sánchez A, Antonioni A. 2021 Gossip and competitive altruism support cooperation in a Public Good game. *Phil. Trans. R. Soc. B* **376**, 20200303. (doi:10.1098/rstb.2020.0303)
91. Samu F, Takács K. 2021 Evaluating mechanisms that could support credible reputations and cooperation: cross-checking and social bonding. *Phil. Trans. R. Soc. B* **376**, 20200302. (doi:10.1098/rstb.2020.0302)
92. Wu J *et al.* 2021 Honesty and dishonesty in gossip strategies: a fitness interdependence analysis. *Phil. Trans. R. Soc. B* **376**, 20200300. (doi:10.1098/rstb.2020.0300)
93. Oishi K, Shimada T, Ito N. 2013 Group formation through indirect reciprocity. *Phys. Rev. E* **87**, 030801. (doi:10.1103/PhysRevE.87.030801)
94. Barrett HC, Saxe RR. 2021 Are some cultures more mind-minded in their moral judgements than others? *Phil. Trans. R. Soc. B* **376**, 20200288. (doi:10.1098/rstb.2020.0288)
95. Radzvilavicius AL, Kessinger TA, Plotkin JB. 2021 Adherence to public institutions that foster cooperation. *Nat. Commun.* **12**, 3567.
96. Chater N, Vitányi P. 2003 Simplicity: a unifying principle in cognitive science? *Trends Cogn. Sci.* **7**, 19–22. (doi:10.1016/S1364-6613(02)00005-0)
97. Van Veelen M, García J, Rand DG, Nowak MA. 2012 Direct reciprocity in structured populations. *Proc. Natl Acad. Sci. USA* **109**, 9929–9934. (doi:10.1073/pnas.1206694109)
98. Abreu D, Rubinstein A. 1988 The structure of Nash equilibrium in repeated games with finite automata. *Econometrica* **56**, 1259–1281. (doi:10.2307/1913097)
99. Binmore KG, Samuelson L. 1992 Evolutionary stability in repeated games played by finite automata. *J. Econ. Theory* **57**, 278–305. (doi:10.1016/0022-0531(92)90037-I)
100. Chatterjee K, Sabourian H. 2009 Game theory and strategic complexity. In *Encyclopedia of complexity and systems science* (ed. R Meyers), pp. 639–658. New York, NY: Springer.
101. Vigo R. 2006 A note on the complexity of Boolean concepts. *J. Math. Psychol.* **50**, 501–510. (doi:10.1016/j.jmp.2006.05.007)
102. Quine WV. 1952 The problem of simplifying truth functions. *Am. Math. Monthly* **59**, 521–531. (doi:10.1080/00029890.1952.11988183)
103. McCluskey EJ. 1956 Minimization of Boolean functions. *Bell Syst. Tech. J.* **35**, 1417–1444. (doi:10.1002/j.1538-7305.1956.tb03835.x)
104. Imhof LA, Fudenberg D, Nowak MA. 2005 Evolutionary cycles of cooperation and defection. *Proc. Natl Acad. Sci. USA* **102**, 10 797–10 800. (doi:10.1073/pnas.0502589102)

105. Fudenberg D, Maskin E. 1990 Evolution and cooperation in noisy repeated games. *Am. Econ. Rev.* **80**, 274–279.
106. Traulsen A, Nowak MA, Pacheco JM. 2006 Stochastic dynamics of invasion and fixation. *Phys. Rev. E* **74**, 011909. (doi:10.1103/PhysRevE.74.011909)
107. Okada I. 2020 A review of theoretical studies on indirect reciprocity. *Games* **11**, 27. (doi:10.3390/g11030027)
108. Brandt H, Ohtsuki H, Iwasa Y, Sigmund K. 2007 A survey of indirect reciprocity. In *Mathematics for ecology and environmental sciences. Biological and medical physics, biomedical engineering* (eds Y Takeuchi, Y Iwasa, K Sato), pp. 21–49. Berlin, Germany: Springer.
109. van Apeldoorn J, Schram A. 2016 Indirect reciprocity; a field experiment. *PLoS ONE* **11**, e0152076. (doi:10.1371/journal.pone.0152076)
110. Vigo R. 2009 Categorical invariance and structural complexity in human concept learning. *J. Math. Psychol.* **53**, 203–221. (doi:10.1016/j.jmp.2009.04.009)
111. Wegener I. 1987 *The complexity of Boolean functions*. New York, NYT: John Wiley & Sons Inc.
112. Garfield ZH, Schacht R, Post ER, Ingram D, Uehling A, MacFarlan SJ. 2021 The content and structure of reputation domains across human societies: a view from the evolutionary social sciences. *Phil. Trans. R. Soc. B* **376**, 20200296. (doi:10.1098/rstb.2020.0296)