PHILOSOPHICAL TRANSACTIONS B

rstb.royalsocietypublishing.org

Review

Cite this article: McAuliffe K, Dunham Y. 2016 Group bias in cooperative norm enforcement. *Phil. Trans. R. Soc. B* **371**: 20150073. http://dx.doi.org/10.1098/rstb.2015.0073

Accepted: 21 September 2015

One contribution of 16 to a theme issue 'Understanding self and other: from origins to disorders'.

Subject Areas:
cognition, behaviour

Keywords:
norm enforcement, group bias, cooperation, fairness, punishment

Author for correspondence:
Katherine McAuliffe
e-mail: katherine.mcauliffe@yale.edu

Group bias in cooperative norm enforcement

Katherine McAuliffe[1,2] and Yarrow Dunham[1]

[1]Department of Psychology, Yale University, 2 Hillhouse Avenue, New Haven, CT 06511, USA
[2]Department of Psychology, Boston College, Chestnut Hill, MA 02467, USA

A hallmark of human social cognition is the tendency for both adults and children to favour members of their own groups. Critically, this in-group bias exerts a strong influence on cooperative decision-making: people (i) tend to share more with members of their in-group and (ii) differentially enforce fairness norms depending on the group membership of their interaction partners. But *why* do people show these group biases in cooperation? One possibility is that the enforcement of cooperative norm violations is an evolved mechanism supporting within-group cooperation (*Norms-Focused Hypothesis*). Alternatively, group bias in cooperation could be a by-product of more general affective preferences for in-group members (*Mere Preferences Hypothesis*). Here, we appraise evidence from studies of both adults and children with the goal of understanding whether one of these two accounts is better supported by existing data. While the pattern of evidence is complex, much of it is broadly consistent with the Mere Preferences Hypothesis and little is uniquely supportive of the Norms-Focused Hypothesis. We highlight possible reasons for this complexity and suggest ways that future work can continue to help us understand the important relationship between group bias and cooperation.

1. Introduction

> A tribe including many members who, from possessing in a high degree the spirit of patriotism, fidelity, obedience, courage, and sympathy, were always ready to aid one another, and to sacrifice themselves for the common good, would be victorious over most other tribes.
>
> —Darwin, *Descent of Man*, Ch. V

Groups are central to social life. Humans are born into a set of ever-widening concentric circles from the family through the local community to the wider society, with each of these circles structuring social interactions by creating a backdrop of norms and behavioural routines that must be mastered [1]. One of the hallmarks of the psychology that arises in these complex contexts is the seemingly pervasive tendency to have positive attitudes—or *preferences*—in favour of social in-groups, i.e. the groups to which one belongs. While this by no means exhausts the nature of intergroup preferences (for instance, status hierarchies also exert powerful influences on social attitudes: [2,3]), it remains one of the most regularly observed phenomena in social psychology. For example, adults across many cultural contexts have in-group preferences with respect to a wide range of social targets, including racial [4], national [5], religious [6] and other ethnic categories (e.g. [7]). The presence of affiliative motives towards such a wide range of social collectives stands in stark contrast to other species, especially our primate cousins, who treat as important a much more limited set of social distinctions, such as biological sex, kinship and dominance status [8]. By contrast, humans are highly flexible and highly promiscuous when it comes to taking part in in-group life.

Critically, in-group biases emerge early in development [2,9] and—in both adults and children—are expressed for previously unfamiliar and randomly assigned social identities, suggesting they do not require protracted enculturation and that the basic cognitive capacities necessary to identify and affiliate with social groups are available from a young age [10–12]. The fact that

THE ROYAL SOCIETY PUBLISHING

in-group bias emerges so early and so pervasively suggests that it is deeply rooted in human psychology, perhaps in the form of a 'preparedness' to identify and affiliate with salient social groups [12,13].

Many thinkers have offered explanations for the group-centric nature of our species. One common refrain with echoes going back to Darwin is that our life in groups is about solving a coordination problem, namely the problem of cooperation. How do I determine whom I can trust? Who is likely to repay costs I incur for their benefit? By specifying a circle of others with whom one might have the opportunity for repeated interaction, groups could serve to identify profitable interaction partners. In developing this kind of account, social norms emerge as the glue that allows within-group interactions to run smoothly: by inculcating members with norms specifying how to behave, social interactions are regularized and rendered predictable. Group members do well by carefully attending to and internalizing group norms, and may further benefit from engaging in norm enforcement, thereby helping to ensure that other group members comply with the norms that help make cooperation profitable (e.g. [14]). Thus, on this account, the individual's focus on and even affiliation with the in-group is functional because it increases commitment to in-group norms that in turn promote cooperation. We refer to this proposal as the *Norms-Focused Hypothesis*. As we will see, there is considerable evidence that is broadly consistent with this proposal, including that individuals preferentially attend to in-group norms and preferentially cooperate with in-group members. However, surprisingly little of this data is *uniquely* supportive of this view, especially when contrasted with potentially simpler and more general accounts of in-group affiliation, which we collectively refer to as the *Mere Preferences Hypothesis*.

This contrasting perspective has its origin in social psychological theories seeking to account for the widespread presence of in-group preference. Their goal is generally to account for in-group preference by elucidating the psychological mechanisms that give rise to it. In this sense, they operate on a more proximate level as compared to the norms-focused view's appeal to ultimate explanations, though they can be readily related to ultimate considerations, a point to which we return below. These views frequently begin by pointing out that social categories, unlike categories in other domains, have a special relationship with the self by virtue of naming an in-group or an out-group. This special relationship implies a special relevance: the groups I belong to are closely linked to me, aspects of who I am and where I reside in the social order. Psychologists have suggested that this simple fact implicates a range of cognitive and affective processes that can explain in-group favouritism. But critically for our purposes, none of them specifically implicate the promotion of cooperation. After all, individuals positively evaluate all sorts of other things that relate to the self, including their abilities, possessions, friends and even trivialities like the letters in their names; in this context, preference for in-groups merely appears to be one part of a broader constellation of self-favouritism (e.g. [15,16]). Indeed, this positive evaluation could, in principle, lead to 'mistakes' in a cooperative context if people systematically overestimate the cooperative tendencies of their group members. Such mistakes would be unlikely to be pay-off-maximizing and so would seem inconsistent with the idea that cooperation and group bias co-evolved.

There are a variety of proposals regarding the precise nature of the relationship between self and group. For example, some argue that emphasizing or even manufacturing positive dimensions of an in-group is a means of enhancing or defending the self (a central tenet of Social Identity Theory [17]); others instead suggest that self-related positivity simply spreads from the self to social in-groups via basic associative processes [18,19]. These details need not concern us here, but the critical point is that once in-group favouritism emerges (via any of these routes), it can directly affect many aspects of intergroup functioning, including how I treat in-group members in cooperative interactions, simply because I prefer in-group members and preferences affect behaviour [20,21]. Thus, these views explain enhanced cooperation with in-group members, but not by postulating that specific norms govern within-group cooperative behaviour. Importantly, as mentioned above, this mere preferences account exists at a different level of analysis, in the form of proximate cognitive-affective explanations for in-group bias rather than ultimate explanations based on the evolved function of groups. To the extent that it does raise evolutionary questions, they seem to be different ones entirely: *why is the self so central? Why self-enhance?* From an evolutionary perspective, self-enhancement has the obvious benefit of ensuring that individuals are invested in their own survival and reproduction, and are persistent in the face of challenge or failure. However, these details are not critical for our purposes here. Rather, we would merely emphasize that self-enhancement is a pervasive human characteristic that is positively associated with in-group bias [18,19]; it thus offers a potential explanation for some forms of cooperative behaviour directed at in-group members. We thus now turn to the question of how these two accounts of the broad tendency to orient towards in-groups differ in what they predict with regard to cooperative norm enforcement.

The question that occupies us here is whether there is in fact a special relationship between groups and cooperative norms. The answer has direct relevance for theories about the evolution and development of cooperative norms as well as norm enforcement more generally; an affirmative answer would suggest that these two aspects of our psychology were subject to the same selective pressures and should be expected to co-occur within a normal range of human behaviour [22]. The difficulty of reaching an answer, however, comes from the fact that most evidence marshalled in favour of the view is equally consistent with the mere preferences view outlined in the preceding paragraph. For example, being more generous towards a novel in-group member could represent the first, norm-compliant move in a repeated cooperative interaction and thus follow from the hypothesized link between groups and norms. But this same behaviour also could represent the simple consequence of preference, i.e. of being more generous towards individuals you like more.

Critically, there is a family of cases that begins to separate the two accounts. These are cases in which preferences and norm compliance begin to point in different directions. Consider cases in which a norm violation is perpetrated by an in-group member. If, as suggested by the norms account, groups are essentially social structures designed to foster cooperation, the violation will be highly salient and objectionable, violating the central compact of group life. This leads to the prediction that it will elicit greater approbation and punishment than if the same norm were violated by an out-group member, who has no special obligation to the group member.

In short, then, the norms-focused account predicts that in-group norm violations will elicit greater punishment from other group members.

The mere preferences account, on the other hand, leads to a quite different prediction. In particular, because in-group members are viewed more positively than out-group members, an in-group violator will be evaluated less negatively and perhaps forgiven more readily than an out-group individual. In short, the mere preferences account suggests that an in-group member's negative action will at least to a degree be offset by the positive evaluation they gain through group membership. Thus, cases of norm violation and norm enforcement are uniquely positioned to weigh in on these two competing accounts: if in-group norm violators are judged *more* harshly than out-group norm violators, the norms-focused account is supported; if in-group norm violators are judged *less* harshly than out-group norm violators, the mere preferences account is supported. Experimental evidence that fits this bill comes primarily from two sources. In the tradition of economic games, second- and third-party costly punishment games have the necessary structure. Consider the Ultimatum Game, a two-player game in which a Player A proposes a split of an endowment between herself and Player B; Player B then can either accept the offer (thereby actualizing the split) or reject the offer (in which case the entire endowment is lost to both players). Because any rejections of non-zero offers by Player B amount to a loss for Player B to impose a loss on Player A, rejections can be conceptualized as costly punishment. In the group setting, an in-group member's low offer might be considered a violation of an in-group norm of fairness, leading to higher motivation to punish (the prediction of the norms-focused account). Conversely, the same offer could be construed as an ambiguous or moderately unfair decision made by a liked other, leading to less motivation to punish (the prediction of the mere preferences account). Thus, rates of rejection in intergroup Ultimatum Games provide one source of evidence concerning the fitness of these two theories.

A second class of cases comes from social psychology in the form of the so-called 'Black Sheep' effect, which has been demonstrated in both adults and children (BSE; [23,24]). In these cases, an in-group member violates a core group norm, and in at least some cases, ends up being disliked *more* than an out-group member who behaves in the same way. For example, supporting the sports team associated with a rival university over one's own university would be a case in which a putative norm of group loyalty is violated, and it would likely lead to greater derogation of that individual as compared to an out-group member who supports either team. To the extent that this pattern emerges, it appears to contradict the prediction of a mere preference account, which would presumably predict greater forgiveness of (otherwise favoured) in-group norm violators. However, it is difficult to assess the BSE's relevance for the topic of group biases in *cooperation* because the effect has generally been demonstrated via violations of specific in-group standards, such as the aforementioned example of supporting a rival sports team. Does the BSE emerge in the same manner when an in-group member fails to cooperate with the participant? We are not aware of BSE studies that explore this question. Thus, the question of whether cooperative norm violations lead to a BSE cannot be answered until we know whether within-group cooperation functions as an in-group standard or not. We view this as an important, but separate, question. Consequently, in what follows, we set aside the BSE and instead focus on studies that directly relate to norms regulating cooperative behaviour.

The key question, then, is whether in-group members are judged more or less harshly than out-group members for violating cooperative norms. But there is an additional factor complicating the assessment of the relevant evidence here, namely that responses to norm violations, such as cases of second- and third-party punishment, are highly variable across cultures (e.g. [25–27]). For example, while punishment of low offers in economic games is common in some cultures, in other cultures so-called 'antisocial' punishment of generous offers frequently occurs [27]. Moreover, the strength of in-group bias, both in terms of cognitive outcomes like stereotyping [28] and behavioural outcomes like resource allocation [29], also differs cross-culturally. That both cooperative norm enforcement and in-group bias show heterogeneity across societies suggests that the extent to which they are related may also vary. Thus, whatever initial relationship holds between groups and norms is most likely shaped in profound ways by cultural experience, raising the possibility that what we observe in adults cannot be taken as a pure reflection of a species-specific orientation towards groups. A key place to look, then, is development, i.e. studies performed with children. Because they have less experience internalizing cultural norms, their attitudes and behaviour can be taken as a more direct window into the initial state of intergroup cognition prior to extensive enculturation. Of course, this point should not be overstated, as culture begins to influence children's development from or even before birth. Nonetheless, children's behaviour frequently differs from that of their cultural elders, offering a suggestive window into the intuitive sociology that guides reasoning about groups [30]. Our goal in this paper, then, is to survey evidence spanning the entire lifespan as a lens through which to understand whether intergroup cognition is grounded in a notion of groups as containers for cooperative norms.

It is important to acknowledge at this point that the two views we are contrasting here, a norms-focused and a mere preferences account of group behaviour, are not always clearly specified or explicitly argued for in this form in the literature. Further, they need not in principle be wholly incompatible. For example, some aspects of the norms-focused approach, such as preferential learning of in-group norms, could be driven by preferences, and in that sense preferences could be a proximate psychological mechanism realizing some aspects of the norms-focused proposal. Despite these complexities, we think there is value in stating the two views clearly and contrasting them to determine whether there is evidence to support the stronger form of the norms-focused approach in which groups are privileged grounds over which cooperative norms are internalized and enforced. Correctly specifying the nature of group effects on cooperation requires understanding whether those effects are a direct consequence of intergroup cognition or merely spill over from in-group positivity more broadly; in both cases, some classes of within-group cooperation might be enhanced, but the reason for that enhancement, and the nature of the relationship between cooperation and parochialism, is quite different across the two accounts. We attempt to do this by reviewing the literature on group effects on cooperation with an eye towards identifying what aspects of the evidence, if any, lend clarity to these two accounts of group effects on cooperative norms.

2. Group bias in sharing behaviour

Both adults and children exhibit striking prosocial tendencies, often sacrificing their own resources to help others. However, humans are not promiscuous givers, and people's generosity is influenced by information about the group membership of social partners. For instance, in the standard Dictator Game, in which one individual is given the unilateral opportunity to offer a proportion of an endowment to another, dictators are more likely to share with members of their in-group than members of their out-group [31–33]. Biases that exist even in these one-shot contexts can shed light on the relationship between cooperation and group bias because they can be taken as the first move in a repeated cooperative interaction, though as we discuss further below, they can also be taken as a simple behavioural consequence of in-group preference.

To elaborate, initial group biases in sharing may result from an expectation of more frequent and beneficial future interactions with in-group members [32,34,35]. Supporting this sort of view, Yamagishi and colleagues found that if a participant's monetary reward is carefully fixed to be completely independent of any action by other participants, the tendency to give more is eliminated, though attitudinal preferences in favour of the in-group remain (summarized in [32]). Other work has shown that when participants receive explicit information that in- and/or out-group members do not display bias in their resource allocation, they show reductions in their own in-group favouritism [10]. A similar effect is observed when participants are informed that their interaction partner will be unaware of their group membership [36]. Together, these findings suggest that at least a portion of in-group bias observed in allocation tasks is likely rooted in participants' expectations of reciprocity: namely, that in-group members will behave more positively towards them than out-group members. Indeed, expectations that in-group members will be reciprocally cooperative appear to boost the effect of in-group favouritism on resource sharing across a range of tasks [37]. Critically, however, real or perceived interdependencies between the players in these games, such as the belief that one's current donation might be reciprocated, are not needed for the emergence of in-group bias in sharing, suggesting that in addition to effects of reciprocity, people simply prefer to give more to in-group members [37], a finding which is compatible with a behavioural manifestation of in-group preference over and above contentful assumptions about within-group reciprocity.

(a) Developmental evidence for group bias in sharing

Given the possibility that norms regulating group conduct become increasingly ingrained across development, stronger evidence regarding parochialism in sharing would come from developmental evidence, and indeed, like adults, children show sensitivity to the identity of social partners. Children attend to group membership, including race, gender and minimal group markers when making allocation decisions [11,38]. They also attend to other group categories like classroom membership. For example, Fehr *et al.* [39] conducted a study with Swiss 3- to 8-year-olds in which participants played three one-shot economic games designed to study prosocial behaviour, sharing and envy. In this task, children were either paired with a member of their own playgroup, class or school (in-group condition) or a member of a different playgroup, class or school (out-group condition). Results showed that children were more likely to be prosocial towards an in-group partner compared to an out-group partner. With increasing age, children were more likely to share with an in-group member, while sharing with out-group members decreased slightly with age. Additionally, boys showed more tolerance of disadvantageous inequality (unequal allocations favouring the partner) if the advantaged partner was from the in-group. Thus, from a relatively young age, children show in-group favouritism in resource decisions.

Studies of children's costly sharing behaviour in unconstrained games like the Dictator Game have also shown that children tend to favour in-group members. For example, Gummerum *et al.* [31] used a minimal groups paradigm to assign German children and adults to groups. Participants then played a Dictator Game with either an in-group or an out-group partner. They found that sixth graders and adults, but not second graders, gave more to in-group members. They additionally found that in-group bias was attenuated in adults when they were given more information about group behaviour, while in-group bias in older children was impervious to additional information. More recently, a study testing 3- to 6-year-olds in Israel built on this paradigm found that in-group bias in the Dictator Game is especially pronounced in boys [40].

In sum, results from sharing tasks conducted with adults and children are broadly consistent with the notion that people see in-group members as reliable cooperation partners, but it is also interpretable as a manifestation of in-group positivity without any rich expectations about cooperation or shared norms. In the next section, we discuss work that more directly tests whether in-group bias and cooperative norms are related in a meaningful way by addressing cases of costly norm enforcement.

3. Does in-group bias influence cooperative norm enforcement?

(a) Do adults show group bias in their enforcement of cooperative norms?

According to the Norms-Focused Hypothesis, people should be more likely to punish norm violations from in- as opposed to out-group members *despite* their in-group bias. To date, work on second- and third-party norm enforcement in cooperative games has provided mixed support for this idea. One example of a study that has provided support for this relationship comes from Shinada *et al.* [41]. They found that individuals who had previously donated in a gift-giving game were more likely to punish in-group than out-group members who had not donated in the same game. This result is consistent with the idea that norm violations are punished more harshly when committed by an in- as opposed to out-group member. A suite of recent studies has also investigated the effects of group bias in decisions in the Ultimatum Game. In one investigation, McLeish & Oxoby [42] primed participants with a collective, school-based identity shared with their game partners, and measured actual offers as well as the minimum acceptable offer in an Ultimatum Game. They found that players offered more to in-group members but also expected more from in-group members, in that they reported higher minimum acceptable offers after being primed with a collective identity. More recently, Mendoza *et al.* [43] conducted an Ultimatum Game in which participants were paired with racial in- and out-group members. They found

that participants were more likely to reject marginally unfair offers ($8 of $20) from in-groups as opposed to out-group members, suggesting that adults are more likely to enforce fairness norm violations within groups.

In contrast to these two studies, Valenzuela & Srivastava [44] found that participants were more tolerant of unfair offers from in-group members. In their game, university students played an Ultimatum Game with a partner from their class (in-group) or students from a competing university (out-group). In the standard version of the Ultimatum Game, in which both players had perfect information about the game's pay-off structure, participants were more likely to accept a marginally unfair offer ($7.50 out of $20.00) if it came from an in-group proposer. Another recent Ultimatum Game study that focused on racial effects also yielded slightly different findings. This study tested whether participants were more likely to reject unfair offers from white versus black proposers [45]. Their findings showed that non-black participants were more tolerant of unfairness from white proposers compared to black proposers, though this effect was not significant when white proposers were analysed separately. Taken together, these four studies of second-party punishment in the context of the Ultimatum Game paint a puzzling picture of the effects of group bias on norm enforcement. They have provided some evidence for the idea that norms are contained within groups and must therefore be enforced therein [41–43] and some evidence in opposition to this idea [44,45].

One potential reason why results from the Ultimatum Game may have generated these conflicting findings is that the game demands that participants resolve a tension between the desire to favour the in-group and the desire to reach a deal that the other party will accept. In support of this idea, a recent study by Stagnaro *et al.* [46] directly contrasted in-group bias in terms of support for pro-life versus pro-choice policies on a Dictator Game versus an Ultimatum Game. While participants favoured their in-group on the Dictator Game, no group bias appeared in proposals in the Ultimatum Game. They suggest that this disconnect appears because, while individuals do hold in-group favouring attitudes, the interdependence of fates in the Ultimatum Game provides a motivation to override bias in order to make offers that are likely to be accepted. In line with this idea, Yamagishi & Kiyonari [47] propose that group bias will not appear in situations that primarily involve direct reciprocity because such exchanges can be resolved by attending directly to the interaction history rather than employing a more general heuristic. By contrast, group bias is expected in situations in which individuating information is absent such that individuals use a heuristic of in-group reciprocity, thereby enabling a system of generalized group benefits. In support of this idea, they find that group bias exists in a simultaneous Prisoner's Dilemma (PD), in which players have no information about what their partner decided, but is absent in a sequential PD, in which players can base their behaviour on their partners last move.

Given that work on second-party enforcement of fairness norms has failed to provide clear evidence for the idea that norms should be preferentially enforced within groups, we now turn to evidence from third-party punishment contexts. These contexts may be more likely to elicit group bias because third-party enforcers have no direct material interest in the exchange that they are observing, eliminating at least one of the motivations alluded to above (the desire to maximize individual profit). Bernhard and co-workers [22] conducted a group-based third-party punishment game with participants in Papua New Guinea. In this game, participants learned about a donation from an in- or out-group dictator to an in- or out-group recipient. They were then given an opportunity to spend money to punish the dictator for his/her donation. Findings from this study showed that participants were more willing to punish selfishness directed at in-group members than out-group members. Contrary to the predictions that norm enforcement should be highest within groups, they did not see preferential punishment of norm violations by in-group members. Rather, they found that punishers were more lenient when norm violations came from in-group dictators (see also [33]). A similar result was seen in a study that randomly assigned participants to real groups (training platoons in the Swiss army) and examined third-party responses to defection in a simultaneous PD game [48]. People were more likely to cooperate with members of their own group, and, as in Bernhard *et al.*, third parties were more likely to punish when an in-group member had been the victim of defection, showing that this finding holds across different economic contexts and group manipulations. Still other studies have shown that third parties are particularly protective of in-group victims when norm violations are committed by out-group members [22,49,50].

Taken together, work on second- and third-party norm enforcement in adults has shown that punishment decisions are importantly influenced by group bias. However, the directionality of these effects appears to fluctuate in second-party contexts, and in third-party contexts the bulk of reported effects are in opposition to the theoretical prediction of the norms-focused view. Given the complexity of results from studies of adult norm enforcement, understanding whether and how children react when confronted with uncooperative behaviour can help shed light on how in-group bias is related to norm enforcement from its inception.

(b) Developmental evidence for bias in norm enforcement

From a young age, children are sensitive to social norms across conventional and moral domains [51–53]. Further, several experimental studies show that children will spontaneously protest the violation of even recently instituted norms [54,55]. For example, German children as young as three who have just learned the rules of a simple game will spontaneously and explicitly correct a puppet that plays the game incorrectly [55]. Critically for our purposes here, recent work has shown that while children enforce moral norm violations with great regularity, they are more selective in enforcing conventional norm violations, and in particular enforce conventional norms more regularly, when the violator is an in-group member [56]. This is consistent with the notion that children take conventional norms to reside at the level of the group, such that an in-group member is uniquely required to conform to them. An important open question that therefore emerges is whether children conceptualize cooperative norms as moral or conventional. If cooperative norms are considered group-based *conventions*, we might expect children to impose them more forcefully on in-group members, while if they are considered general moral obligations, we would expect them to be

imposed more broadly. Unfortunately, the literature does not yet provide a clear answer to this question.

Children's sensitivity to information about group-based norms extends to situations where they are sharing their own resources: children adjust their sharing based on what in-group members have shared [57] and whether they are being observed by an in- or out-group member [58]. Until recently, however, the extent to which children expect or require others to adhere to fairness norms in an intergroup context has remained unknown. More specifically, do children preferentially *enforce* fairness norm violations that have been committed by in-group members, as predicted by the norms-focused account? To our knowledge, only two studies have addressed this question.

First, McAuliffe & Dunham [59] used a minimal group-based Ultimatum Game to test whether children show group bias in second-party fairness norm enforcement. Six- to 10-year-old American children made proposals to in- and out-group members and responded to proposals from in- and out-group members. Findings revealed that children tended to make relatively fair offers and frequently rejected unfair offers. However, despite successfully inducing group bias, the minimal group manipulation had no effect on children's proposals or rejections. As discussed in §2a, the fact that group bias was not observed in the Ultimatum Game may be because this game demands that participants resolve a tension between group loyalty and a desire to reach a mutually beneficial agreement with their partner. Put differently, children are especially reactive to unfairness when it places them in a disadvantageous position [60] and this strong reaction may eclipse group bias effects in games structured like the Ultimatum Game.

Second, Jordan *et al*. [61], tested 6- and 8-year-olds from the USA in a group-based third-party Dictator Game in which they learned about a selfish actor who refused to share with a recipient. Children were assigned to a minimal group based on colour preference (blue and yellow teams) and the group membership of the actors and recipients with regards to the participant were varied such that all grouping combinations were tested within subject. Results showed that both 6- and 8-year-olds were more likely to pay to punish selfish offers from out-group compared to in-group actors. Put another way, children were more forgiving of norm violations committed by in-group actors. Additionally, they showed that 6-year-olds, but not 8-year-olds, were more likely to punish fairness norm violations that negatively affected an in-group member (see [61] for a discussion of this developmental change). These results align with work on group bias in third-party punishment in adults: namely, people appear to be protective of in-group victims and especially punitive of out-group selfishness. Thus, this study did not provide any evidence uniquely compatible with the norms-based view of social groups.

In sum, work on norm enforcement in children is currently rather sparse. However, the little evidence we do have for bias in norm enforcement is more consistent with the mere preferences account than the within-group enforcement account.

4. Discussion

A rich history of work in intergroup psychology shows that humans generally show robust preferences for members of their own groups. Another expansive body of work has explored the ways in which norms govern cooperative behaviour across human societies. Until recently, these lines of research have remained independent. However, over the past decade, researchers have begun to investigate how group bias affects people's behaviour in cooperative contexts, with a specific focus on whether intergroup cognition, including its attendant biases, co-evolved with a norms psychology designed to foster cooperation. Our primary aim in this paper was to survey the empirical evidence in favour of this proposed link between group biases and cooperative norms and to contrast it with a potentially simpler and more general account in which group effects on cooperation stem not from a co-evolved norms psychology but simply via the in-group preferences that routinely follow in the wake of intergroup categorization. While we include a range of resource sharing tasks, as we outlined above, the critical test cases are those in which individuals can pay to enforce fairness norms in in- and out-group contexts.

The lesson from resource sharing tasks is relatively straightforward: people tend to share more with members of the in-group. This pattern has now been repeatedly observed in both adults and young children, suggesting that the tendency to favour group members in resource allocation is deeply ingrained in humans. What is not presently clear, however, is whether bias in sharing tasks is driven by a bias *for* the in-group, *against* the out-group, or both. Future work could help clarify this by always presenting participants with a neutral control (i.e. a choice between an in- or out-group member and a recipient who is not assigned to a group).

In contrast to findings from sharing tasks, results from norm enforcement tasks paint a more complicated picture. A small number of Ultimatum Game studies have provided support for the norm enforcement account by showing that adults are more likely to enforce norm violations committed by in-group members. However, the majority of studies have generated results that are more consistent with the mere preferences account. Namely, people are (i) more likely to enforce norms when in-group members are the victims of a norm violation and (ii) more likely to punish out-group members for violating a norm. These results simply do not conform to the predictions of a norms-focused account in which in-group norm enforcement stabilizes cooperation.

In the developmental arena, there is now clear evidence that children are highly concerned with normativity, even protesting violations of recently created norms, and these protests are particularly targeted at in-group members. However, there is very little developmental work investigating group effects on cooperative norm *enforcement* in children, though quite interestingly, one of the two studies that have been done suggests that both in-group protection and heightened out-group punishment emerge early in development, but preferential enforcement of in-group norms through increased punishment of in-group norm violators does not.

Thus, existing data on in-group bias on cooperative norm enforcement is more consistent with the mere preferences account. However, the conclusions that follow from this should not be overstated. Indeed, it is clear that preferences for the in-group do importantly affect decision-making in the domain of cooperation, for example by leading children and adults to give more of their resources to in-group members. This is perhaps not altogether surprising given that in-group preferences guide behaviour in a range of other domains as well, as well evidenced in the literature on intergroup discrimination [20,21]. However, what we want to

emphasize here is that the evidence we have reviewed suggests some limitations on the nature of the link between groups and cooperative behaviour, and in particular that contra a number of suggestions in the literature, norm enforcement does not appear to be uniquely directed at in-group violators. It is of course possible that in-group preference itself reflects a proximate mechanism designed to foster cooperation, and this is the sense in which existing evidence *is* compatible with a special relationship between in-group bias and cooperation. But this specific relationship would not run via the enforcement of group-specific norms, and indeed, generates a very different set of predictions regarding how groups and norms relate (e.g. that in-group members are more likely to be forgiven rather than punished for norm violation).

(a) Explaining inconsistencies in past work

The research that we have surveyed clearly illustrates that there is dramatic variation in the strength and directionality of the influence of group bias on cooperative norm enforcement. These contexts are intriguing because they raise a tension between group loyalty and a desire to uphold cooperative norms. Previous work has shown that adults show inter-individual variation in how they reconcile the tension between group loyalty and fairness, with individuals differing in the extent to which they weigh these two demands [62]. This finding offers a useful perspective on resolving inconsistencies in past work. First, different cooperative situations or experimental contexts may cause people to value group loyalty over the adherence to cooperative norms (or vice versa). Second, the nature of the 'group' about which individuals are reasoning may also influence how people behave when faced with norm violations.

Past work on group effects on cooperative norm enforcement in both adults and children can broadly be categorized into studies that ask the participant to weigh material interests against fairness considerations when they are directly involved in the interaction (second-party studies; e.g. the Ultimatum Game) and those that allow for the participant to intervene to prevent unfairness between others when they are not directly involved (third-party; e.g. the Third Party Punishment Game). As we detailed above, these two different contexts may importantly influence the extent to which people value group loyalty versus fairness.

Another factor that may explain variation in results from group-based norm enforcement studies is that the nature of the 'group' varies dramatically across studies, ranging from race [45] through school groups [39] and minimal groups [33,48]. Why might this matter? Previous work has shown that minimal group manipulations affect punishment behaviour in different ways than real, culturally salient groups. In particular, participants in minimal groups punished in-group norm violations less harshly than out-group norm violations, but that effect disappeared with real social groups; further, with real but not minimal groups, participants punished more when the victim was an in-group member [63]. Thus, specific properties of the groups may have a large effect on the pattern of results that emerges (though none of these patterns appear in the form of increased punishment directed towards in-group norm violators). Other work has shown that priming similarity versus group identity can differentially affect people's punishment behaviour [64]. In this study, participants showed less tolerance and more punishment of unfairness generated by those they perceived to be similar to themselves, suggesting an egocentric bias in cooperation—i.e. if cooperation is expected from a similar other, uncooperative behaviour is especially egregious. By contrast, people were *more* tolerant of uncooperative behaviour from in-group members, a finding that, like others we have reviewed, is more compatible with a direct effect of group preference. Importantly, no one has yet fully crossed similarity and group membership to identify the relative weight on these two factors. But the suggestion that we would like to make here is that the nature of the group in a given context (including an experimental context) might cue similarity versus other aspects of group membership (such as future interactions) to different degrees. If so, otherwise highly similar studies might produce different effects based on the differential impact of similarity, group identity and the nature of the group identity itself. Thinking carefully about the nature of the group being manipulated in future studies will also shed some light on the question of whether cooperation is specifically an in-group standard. That is, some groups might well have a strong norm of within-group generosity and cooperation, in which case a failure to act in that way might yield less tolerance from in-group members (cf. Black Sheep effect), while other groups might not.

Based on these observations, future work should carefully consider the nature of the group manipulation used to inculcate a group identity, and carefully consider or even manipulate the extent to which this identity implies similarity, and/or future interaction. Further, third-party games, which eliminate one key concern, the desire to maximize individual profit, offer a clearer lens into the role of group-based norm enforcement, and we suggest that future work be focused here.

5. Conclusion

More and more work is beginning to highlight the important connection between in-group bias and cooperation. However, we are in the early days of understanding the shape and origins of this relationship. At present, most existing evidence is consistent with the view that group bias in cooperation exists due to general in-group favouritism. Future work could test this hypothesis by exploring whether a reversal of in-group preferences results in a reversal in behaviour in cooperative contexts. While we currently have little evidence that is uniquely supportive of the Norms-Focused Hypothesis, this may be because we do not yet understand the specific circumstances under which the predictions of this account are met. A push towards refining and standardizing the methodologies used to manipulate cooperative decision-making and group membership will help clarify the complex ways in which group bias affects cooperative norm enforcement.

Authors' contributions. K.M. and Y.D. wrote the article.

Competing interests. We have no competing interests.

Funding. This work was made possible through the generous support of the Greater Good Science Center at the University of California, Berkeley, Florida State University's Philosophy and Science of Self Control Project, and the John Templeton Foundation.

Acknowledgements. We are grateful to Jillian Jordan, Despoina Lioliou and two anonymous reviewers for helpful comments on an earlier version of this manuscript.

References

1. Bronfenbrenner U. 1977 Toward an experimental ecology of human development. *Am. Psychol.* **32**, 513–531. (doi:10.1037/0003-066X.32.7.513)
2. Dunham Y, Chen EE, Banaji MR. 2013 Two signatures of implicit intergroup attitudes: developmental invariance and early enculturation. *Psychol. Sci.* **24**, 860–868. (doi:10.1177/0956797612463081)
3. Mullen B, Brown R, Smith C. 1992 Ingroup bias as a function of salience, relevance, and status: an integration. *Eur. J. Soc. Psychol.* **22**, 103–122. (doi:10.1002/ejsp.2420220202)
4. Nosek BA, Banaji M, Greenwald AG. 2002 Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dyn. Theory Res. Pract.* **6**, 101–115. (doi:10.1037/1089-2699.6.1.101)
5. Mummendey A, Klink A, Brown R. 2001 Nationalism and patriotism: national identification and out-group rejection. *Br. J. Soc. Psychol.* **40**, 159–172. (doi:10.1348/014466601164740)
6. Galen LW, Smith CM, Knapp N. 2011 Perceptions of religious and nonreligious targets: exploring the effects of perceivers' religious fundamentalism. *J. Appl. Soc. Psychol.* **41**, 2123–2143. (doi:10.1111/j.1559-1816.2011.00810.x)
7. Rubinstein Y, Brenner D. 2014 Pride and prejudice: using ethnic-sounding names and inter-ethnic marriages to identify labour market discrimination. *Rev. Econ. Stud.* **81**, 389–425. (doi:10.1093/restud/rdt031)
8. Wilson ML, Wrangham RW. 2003 Intergroup relations in chimpanzees. *Annu. Rev. Anthropol.* **32**, 363–392. (doi:10.1146/annurev.anthro.32.061002.120046)
9. Aboud FE. 1988 *Children and prejudice*. New York, NY: Blackwell Publishing.
10. Locksley A, Ortiz V, Hepburn C. 1980 Social categorization and discriminatory behavior: extinguishing the minimal intergroup discrimination effect. *J. Pers. Soc. Psychol.* **39**, 773–783. (doi:10.1037/0022-3514.39.5.773)
11. Dunham Y, Baron AS, Carey S. 2011 Consequences of 'minimal' group affiliations in children. *Child Dev.* **82**, 793–811. (doi:10.1111/j.1467-8624.2011.01577.x)
12. Nesdale D, Flesser D. 2001 Social identity and the development of children's group attitudes. *Child Dev.* **72**, 506–517. (doi:10.1111/1467-8624.00293)
13. Dunham Y. 2011 An angry = outgroup effect. *J. Exp. Soc. Psychol.* **47**, 668–671. (doi:10.1016/j.jesp.2011.01.003)
14. Bowles S. 2006 Group competition, reproductive leveling, and the evolution of human altruism. *Science* **314**, 1569–1572. (doi:10.1126/science.1134829)
15. Greenwald AG, Banaji MR. 1995 Implicit social cognition—attitudes, self-esteem, and stereotypes. *Psychol. Rev.* **102**, 4–27. (doi:10.1037/0033-295X.102.1.4)
16. Leary MR. 2007 Motivational and emotional aspects of the self. *Annu. Rev. Psychol.* **58**, 317–344. (doi:10.1146/annurev.psych.58.110405.085658)
17. Tajfel H, Turner JC. 1986 The social identity theory of inter-group behavior. In *Psychology of intergroup relations* (eds S Worchel, LW Austin), pp. 7–24. Chicago, IL: Nelson-Hall.
18. Greenwald AG, Banaji MR, Rudman LA, Farnham SD, Nosek BA, Mellott DS. 2002 A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychol. Rev.* **109**, 3–25. (doi:10.1037/0033-295X.109.1.3)
19. Gramzow RH, Gaertner L. 2005 Self-esteem and favoritism toward novel in-groups: the self as an evaluative base. *J. Pers. Soc. Psychol.* **88**, 801–815. (doi:10.1037/0022-3514.88.5.801)
20. Schütz H, Six B. 1996 How strong is the relationship between prejudice and discrimination? A meta-analytic answer. *Int. J. Intercult. Relat.* **20**, 441–462. (doi:10.1016/0147-1767(96)00028-4)
21. Greenwald AG, Poehlman TA, Uhlmann EL, Banaji MR. 2009 Understanding and using the implicit association test: III. Meta-analysis of predictive validity. *J. Pers. Soc. Psychol.* **97**, 17–41. (doi:10.1037/a0015575)
22. Bernhard H, Fischbacher U, Fehr E. 2006 Parochial altruism in humans. *Nature* **442**, 912–915. (doi:10.1038/nature04981)
23. Marques JM, Paez D. 1994 The 'black sheep effect': social categorization, rejection of ingroup deviates, and perception of group variability. *Eur. Rev. Soc. Psychol.* **5**, 37–68. (doi:10.1080/14792779543000011)
24. Abrams D, Palmer SB, Rutland A, Cameron L, Van de Vyver J. 2014 Evaluations of and reasoning about normative and deviant ingroup and outgroup members: development of the black sheep effect. *Dev. Psychol.* **50**, 258–270. (doi:10.1037/a0032461)
25. Henrich J *et al.* 2006 Costly punishment across human societies. *Science* **312**, 1767–1770. (doi:10.1126/science.1127333)
26. Henrich J *et al.* 2010 Markets, religion, community size, and the evolution of fairness and punishment. *Science* **327**, 1480–1484. (doi:10.1126/science.1182238)
27. Herrmann B, Thoni C, Gächter S. 2008 Antisocial punishment across societies. *Science* **319**, 1362–1367. (doi:10.1126/science.1153808)
28. Spencer-Rodgers J, Williams MJ, Hamilton DL, Peng K, Wang L. 2007 Culture and group perception: dispositional and stereotypic inferences about novel and national groups. *J. Pers. Soc. Psychol.* **93**, 525–543. (doi:10.1037/0022-3514.93.4.525)
29. Hruschka D *et al.* 2014 Impartial institutions, pathogen stress and the expanding social network. *Hum. Nat.* **25**, 567–579. (doi:10.1007/s12110-014-9217-0)
30. Rhodes M. 2012 Naïve theories of social groups. *Child Dev.* **83**, 1900–1916. (doi:10.1111/j.1467-8624.2012.01835.x)
31. Gummerum M, Takezawa M, Keller M. 2009 The influence of social category and reciprocity on adults' and children's altruistic behavior. *Evol. Psychol.* **7**, 295–316. (doi:10.1177/147470490900700212)
32. Yamagishi T, Jin N, Kiyonari T. 1999 Bounded generalized reciprocity: ingroup boasting and ingroup favoritism. *Adv. Group Process.* **16**, 161–197.
33. Chen Y, Li SX. 2009 Group identity and social preferences. *Am. Econ. Rev.* **99**, 431–457. (doi:10.1257/aer.99.1.431)
34. Gaertner L, Insko CA. 2000 Intergroup discrimination in the minimal group paradigm: categorization, reciprocation, or fear? *J. Pers. Soc. Psychol.* **79**, 77–94. (doi:10.1037/0022-3514.79.1.77)
35. Rabbie JM, Schot JC, Visser L. 1989 Social identity theory: a conceptual and empirical critique from the perspective of a behavioural interaction model. *Eur. J. Soc. Psychol.* **19**, 171–202. (doi:10.1002/ejsp.2420190302)
36. Ockenfels A, Werner P. 2014 Beliefs and ingroup favoritism. *J. Econ. Behav. Organ.* **108**, 453–462. (doi:10.1016/j.jebo.2013.12.003)
37. Balliet D, Wu J, De Dreu CKW. 2014 Ingroup favoritism in cooperation: a meta-analysis. *Psychol. Bull.* **140**, 1556–1581. (doi:10.1037/a0037737)
38. Renno MP, Shutts K. 2015 Children's social category-based giving and its correlates: expectations and preferences. *Dev. Psychol.* **51**, 533–543. (doi:10.1037/a0038819)
39. Fehr E, Bernhard H, Rockenbach B. 2008 Egalitarianism in young children. *Nature* **454**, 1079–1083. (doi:10.1038/nature07155)
40. Benozio A, Diesendruck G. 2015 Parochialism in preschoolers' resource distribution. *Evol. Hum. Behav.* **36**, 256–264. (doi:10.1016/j.evolhumbehav.2014.12.002)
41. Shinada M, Yamagishi T, Ohmura Y. 2004 False friends are worse than bitter enemies. *Evol. Hum. Behav.* **25**, 379–393. (doi:10.1016/j.evolhumbehav.2004.08.001)
42. McLeish KN, Oxoby RJ. 2011 Social interactions and the salience of social identity. *J. Econ. Psychol.* **32**, 172–178. (doi:10.1016/j.joep.2010.11.003)
43. Mendoza SA, Lane SP, Amodio DM. 2014 For members only: ingroup punishment of fairness norm violations in the ultimatum game. *Soc. Psychol. Person. Sci.* **5**, 662–670. (doi:10.1177/1948550614527115)
44. Valenzuela A, Srivastava J. 2012 Role of information asymmetry and situational salience in reducing intergroup bias: the case of ultimatum games. *Person. Soc. Psychol. Bull.* **38**, 1671–1683. (doi:10.1177/0146167212458327)
45. Kubota JT, Li J, Bar-David E, Banaji MR, Phelps EA. 2013 The price of racial bias: intergroup negotiations in the ultimatum game. *Psychol. Sci.* **24**, 2498–2504. (doi:10.1177/0956797613496435)

46. Stagnaro N, Dunham Y, Rand DG. In preparation. Self-interest versus in-group bias: inhibiting group favoritism in economic games.
47. Yamagishi T, Kiyonari T. 2000 The group as the container of generalized reciprocity. *Soc. Psychol. Q* **63**, 116–132. (doi:10.2307/2695887)
48. Goette L, Huffman D, Meier S. 2012 The impact of social ties on group interactions: evidence from minimal groups and randomly assigned real groups. *Am. Econ. J. Microecon.* **4**, 101–115. (doi:10.1257/mic.4.1.101)
49. Baumgartner T, Götte L, Gügler R, Fehr E. 2011 The mentalizing network orchestrates the impact of parochial altruism on social norm enforcement. *Hum. Brain Mapp*. **33**, 1452–1469. (doi:10.1002/hbm.21298)
50. Schiller B, Baumgartner T, Knoch D. 2014 Intergroup bias in third-party punishment stems from both ingroup favoritism and outgroup discrimination. *Evol. Hum. Behav.* **35**, 169–175. (doi:10.1016/j.evolhumbehav.2013.12.006)
51. Turiel E. 1983 *The development of social knowledge: morality and convention*. Cambridge, UK: Cambridge University Press.
52. Killen M, Smetana JG. 2005 *Handbook of moral development.* Manwah, NJ: Lawrence Erlbaum Associates.
53. Smetana JG. 1984 Toddlers' social interactions regarding moral and conventional transgressions. *Child Dev.* **55**, 1767–1776. (doi:10.2307/1129924)
54. Rakoczy H, Schmidt MF. 2012 The early ontogeny of social norms. *Child Dev. Perspect.* **7**, 17–21. (doi:10.1111/cdep.12010)
55. Rakoczy H, Warneken F, Tomasello M. 2008 The sources of normativity: young children's awareness of the normative structure of games. *Dev. Psychol.* **44**, 875–881. (doi:10.1037/0012-1649.44.3.875)
56. Schmidt MFH, Rakoczy H, Tomasello M. 2012 Young children enforce social norms selectively depending on the violator's group affiliation. *Cognition* **124**, 325–333. (doi:10.1016/j.cognition.2012.06.004)
57. Chiang Y-S, Wu C-I. 2015 Social influence and the adaptation of parochial altruism: a dictator-game experiment on children and adolescents under peer influence. *Evol. Human Behav*. **36**, 430–437. (doi:10.1016/j.evolhumbehav.2015.03.007)
58. Engelmann JM, Over H, Herrmann E, Tomasello M. 2013 Young children care more about their reputation with ingroup members and potential reciprocators. *Dev. Sci.* **16**, 952–958. (doi:10.1111/desc.12086)
59. McAuliffe K, Dunham Y. In preparation. Fairness overrides group bias in children's second-party punishment.
60. Blake PR, McAuliffe K. 2011 I had so much it didn't seem fair: eight-year-olds reject two forms of inequity. *Cognition* **120**, 215–224. (doi:10.1016/j.cognition.2011.04.006)
61. Jordan JJ, McAuliffe K, Warneken F. 2014 Development of in-group favoritism in children's third-party punishment of selfishness. *Proc. Natl Acad. Sci. USA* **111**, 12 710–12 715. (doi:10.1073/pnas.1402280111)
62. Waytz A, Dungan J, Young L. 2013 The whistleblower's dilemma and the fairness-loyalty tradeoff. *J. Exp. Soc. Psychol.* **49**, 1027–1033. (doi:10.1016/j.jesp.2013.07.002)
63. Goette L, Huffman D, Meier S. 2006 The impact of group membership on cooperation and norm enforcement: evidence using random assignment to real social groups. *Am. Econ. Rev.* **96**, 212–216. (doi:10.1257/000282806777211658)
64. Mussweiler T, Ockenfels A. 2013 Similarity increases altruistic punishment in humans. *Proc. Natl Acad. Sci. USA* **110**, 19 318–19 323. (doi:10.1073/pnas.1215443110)