Cognition 153 (2016) 43–51

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Original Articles

Preschoolers value those who sanction non-cooperators

Amrisha Vaish [a,*], Esther Herrmann [b], Christiane Markmann [b], Michael Tomasello [b]

[a] Department of Psychology, University of Virginia, United States
[b] Department of Developmental and Comparative Psychology, Max Planck Institute for Evolutionary Anthropology, Germany

ARTICLE INFO

Article history:
Received 29 August 2015
Revised 8 April 2016
Accepted 17 April 2016
Available online 29 April 2016

Keywords:
Norm enforcement
Second-order cooperation
Punishment
Reputation
Moral development

ABSTRACT

Large-scale human cooperation among unrelated individuals requires the enforcement of social norms. However, such enforcement poses a problem because non-enforcers can free ride on others' costly and risky enforcement. One solution is that enforcers receive benefits relative to non-enforcers. Here we show that this solution becomes functional during the preschool years: 5-year-old (but not 4-year-old) children judged enforcers of norms more positively, preferred enforcers, and distributed more resources to enforcers than to non-enforcers. The ability to sustain not only first-order but also second-order cooperation thus emerges quite early in human ontogeny, providing a viable solution to the problem of higher-order cooperation.

1. Introduction

Humans regularly cooperate with others, often even with strangers and often even at a cost to themselves (Sober & Wilson, 1998). Since such cooperation results in a greater loss for the cooperating individuals than for free riders (who benefit from the outcomes of the cooperation without investing any resources), it is a puzzle how such cooperation could evolve and be maintained. The classic theories of kin selection and reciprocity provide some answers, but they cannot explain cooperation in large groups of unrelated individuals (Sripada, 2005). One effective solution to the puzzle of large-scale cooperation is that those who break the norms of cooperation are punished, which induces the norm-violators to cooperate more in future interactions and thus enforces the norms of cooperation (Boyd & Richerson, 1985; Nowak, 2006; Nowak & Sigmund, 2005).

However, norm enforcement can be costly and risky to the enforcer. Despite these costs, people across numerous cultures are willing to pay costs to punish non-cooperators and thus enforce cooperative norms (Fehr & Gächter, 2002; Henrich, 2004). Such norm enforcement can itself be considered a cooperative act because in addition to the enforcer, all other members of the group also benefit from the non-cooperator's increased future cooperation (Yamagishi, 1986). A second-order problem of cooperation thus arises: If enforcers pay costs and take risks to enforce norms on non-cooperators, but the non-cooperator's increased future cooperation benefits not only the enforcer but also other group members, then enforcers are at a disadvantage relative to non-enforcers. How, then, can the costly and risky enforcement of cooperative norms evolve and be maintained?

One possibility is that enforcers receive benefits for their punitive behavior that non-enforcers do not receive (Barclay, 2006; Fessler & Haley, 2003; Gintis, Smith, & Bowles, 2001). For instance, enforcers may be seen to be more committed to the group and its norms, less willing to tolerate norm violations, and more trustworthy than non-enforcers. Enforcers may thus be judged more positively, respected, preferred, and more likely to be selected as cooperative partners than non-enforcers (Fessler & Haley, 2003; Frank, 1988). Moreover, as norm enforcement can be considered a cooperative act, and as cooperative people receive more material rewards from group members than less cooperative people (e.g., Milinski, Semmann, & Krambeck, 2002; Wedekind & Milinski, 2000), enforcers may also receive more material rewards than non-enforcers.

A few empirical studies have examined the question of how costly norm enforcement could be sustained (e.g., Barclay, 2006; Horita, 2010; Kiyonari & Barclay, 2008; Nelissen, 2008) and have shown that enforcers do typically receive more reputational and material benefits than non-enforcers (though these effects are not unequivocal and adults may even disapprove of particularly severe or aggressive norm enforcement; Eriksson, Andersson, & Strimling, 2016). However, these studies have all involved adults, leaving unclear when in ontogeny this solution to the problem of second-order cooperation becomes functional. In other words, we

* Corresponding author at: Department of Psychology, University of Virginia, 485 McCormick Road, Gilmer Hall, Room 102, Charlottesville, VA 22903, United States.
E-mail address: vaish@virginia.edu (A. Vaish).

http://dx.doi.org/10.1016/j.cognition.2016.04.011
0010-0277/

do not yet know whether and when children begin to contribute to maintaining norm enforcement and cooperation in the sophisticated ways that are required for large-scale human cooperation. To draw this conclusion, one must study young children's evaluative judgments of enforcers and non-enforcers.

There is a rapidly growing body of developmental work on children's evaluations of first-order norm violations. This work shows that by 3–5 years of age, children protest against first-order transgressions and punish, avoid helping, and distribute fewer resources to first-order transgressors (e.g., Kenward & Dahl, 2011; Kenward & Östh, 2012; Kenward & Östh, 2015; Riedl, Jensen, Call, & Tomasello, 2015; Salali, Juda, & Henrich, 2015; Smetana, Schlagman, & Adams, 1993; Vaish, Carpenter, & Tomasello, 2010; Vaish, Missana, & Tomasello, 2011). In contrast, research on children's responses to second-order norm violations is very sparse. We are aware of only two studies that have broached this question. In one study, 8-month-old infants were shown to prefer to touch a "taker" puppet that had taken a toy away from an antisocial puppet rather than a "giver" puppet that had given a toy to an antisocial puppet (Hamlin, Wynn, Bloom, & Mahajan, 2011). This result was interpreted as showing that infants prefer characters who act negatively towards (or punish) antisocial others. However, as the study was conducted with young infants and used the rather non-specific measure of touching, it is unclear what the nature of infants' evaluations was. For instance, rather than evaluating the taker as a punisher of the antisocial character, perhaps infants preferred the actor who behaved in line with their own evaluations (i.e., behaved negatively towards the antisocial character); indeed, the study's authors themselves acknowledge a similar alternative interpretation (Hamlin et al., 2011). Without more differentiated measures of children's evaluations and some insight into the reasoning behind the evaluations, it is difficult to know whether the mechanisms that sustain second-order cooperation are indeed present in childhood.

A second study examined whether 4-year-old children identified more with (in the sense of choosing to re-enact the role of) a punisher of first-order transgressors than a non-punisher (Kenward & Östh, 2012). The study revealed that although children approved of punishing first-order transgressors, they did not identify more with punishers than non-punishers, hinting that by 4 years of age, children may not yet value norm enforcers. However, as the main focus of that study was not on children's evaluations of enforcers versus non-enforcers, Kenward and Östh did not examine this question systematically or in detail. We thus currently know very little about children's responses to second-order cooperation.

In the present study, therefore, we presented 4- and 5-year-old children with scenarios in which transgressors broke a moral norm by causing harm to a victim. The norm of not causing harm was then either enforced by a norm enforcer, or was not enforced by a non-enforcer. After watching these scenarios, children were asked to evaluate the enforcer and non-enforcer and their behavior, and children's personal preferences for the enforcer versus non-enforcer were assessed. Finally, children were given the opportunity to distribute flowers between the enforcer and non-enforcer in order to assess whether they would provide more resources to the enforcer than the non-enforcer.

The decision to test 4- and 5-year-olds was guided by relevant prior research in which children of similar ages were successfully tested using a similar procedure and which was also concerned with children's understanding of relatively complex cooperation and group norms (Misch, Over, & Carpenter, 2014; Vaish, Carpenter, & Tomasello, 2011). In those studies, 5-year-olds evaluated positively, preferred, and distributed more resources to (a) a moral transgressor who displayed remorse more than one who displayed no remorse (Vaish et al., 2011), and (b) a loyal group member more than a disloyal one (Misch et al., 2014), whereas 4-year-olds did not. Because the present study was also concerned with children's emerging understanding of rather sophisticated norms of cooperation and because our method was adapted from these prior studies, we expected that 5-year-old children should evaluate positively, prefer, and distribute more resources to enforcers than non-enforcers, whereas 4-year-olds may not yet show these effects (as also hinted at by the results of Kenward and Östh (2012)).

2. Method

2.1. Participants

Participants were 4-year-old children (N = 24, 12 girls) between 54 months, 6 days and 59 months, 9 days (M = 56 months, 2 days; SD = 1 month, 20 days) and 5-year-old children (N = 24, 12 girls) between 66 months, 6 days and 71 months, 14 days (M = 68 months, 16 days; SD = 1 month, 22 days). Five additional children were tested but excluded due to experimenter error (n = 2 4-year-olds) or unwillingness to participate (n = 2 4-year-olds and n = 1 5-year-old). All children were native German speakers whose parents had given permission for them to participate in child development studies. Children were recruited from and tested in their daycare centers in a medium-sized German city.

2.2. Design and materials

During the experiment, children sat at a table on which two identical laptop computers were placed next to one another, one to the left and one to the right of the child. All videos were played using the full-screen option in Quicktime Player. A camera recorded a frontal view of the children and a microphone placed between the computers supplied sound to the camera. The procedure had two phases. In each phase, children saw one Enforcement and one Non-enforcement video, about which they received comprehension probe questions (as manipulation checks, i.e., to make sure they grasped the content of the videos) and eight test questions. After the second phase (with a second set of Enforcement and Non-enforcement videos), children received a distribution of resources task and one final test question about why they had distributed the resources in the way that they had. Thus, altogether, children watched four videos (two per phase) and answered 17 test questions (eight after each of the two phases and one after the distribution of resources task).

2.3. Video stimuli

Videos featuring three adult actresses (research assistants in the lab) served as stimuli. Each video featured a 'transgressor' intentionally harming a 'victim,' i.e., breaking the moral norm that one ought not to cause intentional harm to innocent others. An 'observer' watched the interaction, expressed disapproval of the transgression, and then either enforced the moral norm on the transgressor (Enforcement video) or did not enforce the norm (Non-enforcement video). The roles of transgressor (Lisa) and victim (Anya) were always played by the same actresses in all videos, while two different actresses (Susie and Tina) played both the enforcer and the non-enforcer roles across the videos. Each video featured one target object: a doll, ball, clay bird, or picture.

All videos began with three actresses seated around a table: the victim, the transgressor, and one of the two observers – either the enforcer or the non-enforcer. Anya (the future victim, sitting on the left) excitedly brought out and presented the target object to Lisa (the transgressor, sitting on the right) for approximately 15 s, as follows:

Doll. Anya said this was her doll and then happily showed off the doll's hair, eyes, etc.
Ball. Anya said this was her ball and then happily described how colorful it was, the patterns on it, etc.
Bird. Anya brought out a clay bird that she said was hers and announced that she wanted to finish working on it. She then added a crest to the bird's head with some more clay, after which she proudly announced that she had made the bird and stated how pretty it was.
Picture. Anya brought out a drawing of a butterfly that she said was hers and announced that she wanted to finish working on it. She then proudly completed the drawing by adding the antennae, after which she happily stated that she had drawn the picture and that it was very pretty.

At the end of each presentation, Lisa grabbed the object out of Anya's hand, looked at it neutrally for about 5 s, and then announced that she would destroy it. She then proceeded to break off the doll's head, tear out a piece of the ball and pull out some stuffing, break off the wings and the crest of the bird, or tear the picture into four pieces, each in a mildly aggressive way. Lisa placed the target object and the broken pieces on the table. Her actions took a total of approximately 5 s, during which both Anya and the observer (seated in the middle) watched her, with Anya looking upset and the observer looking neutral.

As soon as Lisa had placed the object and pieces on the table, she got up and left the scene (i.e., was no longer visible in the video). Once she was gone, the observer commented neutrally to herself: "Lisa broke Anja's [target object]. I don't think that was good." Thus, both the enforcer and non-enforcer privately expressed their disapproval, ensuring that they both demonstrated similar and reasonable reactions to the transgressions but that their disapproval could not be perceived as enforcement of the norm upon the transgressor.

The transgressor then returned to her seat, at which point the observer either reacted by enforcing or not enforcing the norm on the transgressor. In the Enforcement videos, the observer leaned forward towards Lisa and said to her in a displeased manner, "Hey, you've broken Anya's [target object]! You aren't allowed to do that. Don't ever do that again" (see Fig. 1). In the Non-enforcement videos, the observer did not lean forward and instead said to Lisa in a neutral manner, "Oh, you've broken Anya's [target object]. Hmm, the [target object]. Now it's lying here on the table."

While the observer spoke, Anya continued to look sadly at the broken objects, and Lisa, still looking somewhat aggressive, looked at the observer when the observer first began speaking and then looked back at the broken objects. The observer also looked at the broken objects as she finished speaking. The video ended with a still frame of this scene, which remained on the screen for 6 s. The duration of each video was approximately 1 min.

2.4. Counterbalancing

For each target object, we created four videos in which each observer (Susie and Tina) played the enforcer and the non-enforcer roles. For example, there were four videos of the doll situation: Susie as enforcer, Susie as non-enforcer, Tina as enforcer, and Tina as non-enforcer. There were thus 16 videos in all (four per target object), although each child only watched four of the 16 videos (one per target object). During testing, the doll and ball videos (both featuring the victim's possessions) were always presented together, as were the bird and picture videos (both featuring objects that the victim had created).

Children were randomly assigned to one of 24 presentation orders which counterbalanced which observer was the enforcer versus non-enforcer, whether an Enforcement or Non-enforcement video was presented first, whether the video on the left or the right computer was presented first, and whether children saw the bird and picture or the ball and doll video pair first. Other factors that were counterbalanced will be mentioned below.

Fig. 1. A still frame from an Enforcement video, showing the enforcer, in the middle, chiding the transgressor, on the right, for destroying the clay bird.

2.5. Procedure

All children were tested by the same female adult experimenter (E), who always sat to their left during the experiment (and who was not featured in the videos). E told children that she was going to show them videos of some people doing some things, and that they should watch carefully and then she would ask them some questions. E opened the first video assigned to the child, introduced the three characters on the still opening frame, and played the video (e.g., an Enforcement video of the ball situation on the left computer).

At the end of the video, E paused the still frame and asked the first comprehension probe: "So, Lisa broke Anya's [target object]. Did Susie/Tina think that was good or not good?" The child was expected to answer, "Not good" or something similar. This first probe was to ensure that the child understood that the observer did not approve of the transgression. Only once the child's response indicated this, E moved on and said, "That's right," and asked the second comprehension probe: "And did she [pointing to observer] say that Lisa should never do that again or did she say that the [target object] is lying on the table?" (Order of "Lisa should never do that again" and "the [target object] is lying on the table" was counterbalanced across children). This second probe was to ensure that the child grasped whether or not the observer had enforced the norm, which was critical if the child was going to draw any inferences on this basis. If the child answered correctly ("Should not do that again" or something similar in the Enforcement case; "Lying on the table" or something similar in the Non-enforcement case), E said, "That's right. You've understood it correctly. Let's watch that film again." E then replayed the video and again paused it on the final still frame. If, however, the child answered the second probe incorrectly (e.g., "Should not do that again" in the Non-enforcement case), E said, "Hmm, I'm not so sure about that. Let's watch the film again and I'll ask you the questions again afterwards." E then replayed the video, paused it on the still frame, and repeated both comprehension probes as before. If the child still answered the second probe incorrectly, E corrected her by saying, "No, Lisa broke Anya's [target object], remember? And

Susie/Tina said that Lisa shouldn't do that again/that the [target object] is broken and lying on the table." (Thus, regardless of whether or not children correctly answered the comprehension probes, all children saw each video twice.)

E then opened the second video, which was in the other condition and on the other computer (e.g., a Non-enforcement video of the doll situation on the right computer). She reminded the child of the characters' names, and then followed the same procedure as with the first video. Finally, after the child had seen both videos and answered the comprehension probes, E provided a reminder, for example: "So, here [pointing to first computer screen], Lisa broke Anya's ball and Susie/Tina didn't think that was good. Then Susie/Tina told Lisa that she should never do that again. And here [pointing to second computer screen], Lisa broke Anya's doll and Tina/Susie didn't think that was good. Then Tina/Susie said that the doll is lying on the table" (always starting with the first video children had seen in that phase). While providing this reminder (and throughout the procedure), E was careful to speak neutrally and not to nod or shake her head or in any other way provide evaluations about the scenarios. E then asked the following test questions:

1. Acted rightly: "Which of the two did the right thing? – Susie or Tina?" (pointing to each in turn)

1a. Acted rightly-justification: "Why was it right?"

2. Child plays: "Whom would you prefer to play with? – Susie or Tina?" (pointing to each)

2a. Child plays-justification: "Why would you like to play with her more?"

3. Child dislikes: "Whom do you not like much? – Susie or Tina?" (pointing to each)

3a. Child dislikes-justification: "Why do you not like her much?"

4. Not good: "Which of the two is not good? – Susie or Tina?" (pointing to each)

4a. Not good-justification: "Why is she not good?"

Note that questions 1 and 4 concerned children's judgments of the observers and their behavior, whereas questions 2 and 3 concerned children's personal preferences for one or the other observer. The questions were forced-choice questions because our aim was to assess whether, when presented with the choice, children would be able to use the information about the observers' responses to answer in the hypothesized ways. In response to these forced-choice questions, children were expected to name and/or point to one observer. If a child responded "Both" or "Neither," E prompted her to choose one. If a child did not respond within 5–7 s, E repeated the question once, but if the child still did not respond, E moved on to the next question. Questions 1a, 2a, 3a, and 4a were designed to elicit justifications for children's responses to the forced-choice questions. E thus let children respond freely to these questions and did not probe further. Note that E did not provide any feedback on the correctness of children's responses for any of the test questions. Following these eight test questions, E repeated the entire procedure and all of the questions with the second pair of videos (Phase 2).

For a given child, Phases 1 and 2 were matched in terms of which observer was the enforcer versus non-enforcer, the order in which the observers' names appeared in the test questions, which computer (left or right) the enforcer and non-enforcer appeared on, and which computer (left or right) the first video of the pair was played on. However, the order of the test questions was counterbalanced across children and across the two phases for a given child. For instance, one child received the test questions in the order 1-2-3-4 in Phase 1 and 4-3-2-1 in Phase 2, a different child received the order 1-2-4-3 in phase 1 and 4-3-1-2 in phase 2, and so on.

Finally, after the second phase, children took part in the distribution of resources task. E said that she would see Susie and Tina soon and could bring them something from the child. Then, in front of each computer, E placed a small container holding a photograph of the observer featured on the corresponding computer (the photographs featured the observers looking neutrally at the camera). E then gave the child three cloth flowers to distribute as she wanted. If the child did not distribute all the flowers or asked E for guidance, E encouraged her to decide for herself.

2.6. Coding and reliability

The primary coder first transcribed all children's verbal and/or pointing responses. From these transcriptions, she coded whether children responded correctly to the comprehension probes. Since E asked the second comprehension probe only after children answered the first probe satisfactorily, coding of responses to the first probe was only to make sure that E had followed this procedure and thus that all children understood that the observers disapproved of the transgression. Coding of responses to the second probe assessed whether children grasped the enforcement or non-enforcement right away, whether they grasped it after watching the video again, or whether E provided them the information. If a child did not provide a response, she did not receive a code for that particular question. For reliability, a second coder coded the responses of a random 25% of the sample ($n = 6$ in each age group). Reliability was perfect, $\kappa = 1$.

The primary coder also coded children's responses to the forced-choice test questions (questions 1, 2, 3, and 4) from her transcriptions. Responses were scored '1' if they were consistent with the hypotheses that children should (1) judge that the enforcer did the right thing, (2) themselves prefer to interact (play) with the enforcer, (3) express a dislike for the non-enforcer, and (4) judge the non-enforcer to not be a good person; responses not consistent with these hypotheses were scored '0.' If a child did not provide a relevant response (indicating either the enforcer or non-enforcer), she did not receive a score for that particular question. A second coder coded the responses for a random 25% of the sample. Again, reliability was perfect, $\kappa = 1$.

Children's distribution of the three flowers was coded from videotape and scored 0, 1, 2, or 3 to represent how many flowers children gave to the enforcer. A second coder coded this for a random 25% of the sample. Reliability was perfect, $\kappa = 1$.

Finally, children's justifications (i.e., their responses to test questions 1a, 2a, 3a, and 4a) were coded from transcriptions and assigned scores of either 1 or 0 (see Table 1 for details of the coding scheme). A score of 1 was assigned to justifications that indicated relevant and sophisticated reasoning about the observers and their responses. These included justifications that referred to the norm enforcement or the lack thereof, or involved moral evaluations. Note that references to the enforcement (or the lack thereof) could be of two kinds: One kind – 'Enforcement (repeated)' – involved repeating phrases that had been used in the videos or by E (e.g., "Because she said, 'Don't ever do that again.'"), whereas the other kind – 'Enforcement (redescribed)' – involved using phrases other than those used in the videos or by E (e.g., "Because she didn't care that [the transgressor] tore up the picture."). Although 'Enforcement (repeated)' was clearly relevant to our question of how children perceive norm enforcers, when children repeated the word or phrase used in the videos or by E, it was difficult to know whether they were engaging in higher-level reasoning or not. That is, did they in fact understand the observers' reactions in a sophisticated way but and expressed this by repeating the words they heard before, or did they not understand the observers' reactions in a

Table 1
Coding scheme for justifications.

Score	Category	Content
1	Enforcement	Observer did (or did not) enforce the norm; e.g., "Because she did the right thing and really scolded her but she [non-enforcer] didn't;" "Because you should scold people who break things;" "Because she didn't scold her properly"
	Enforcement (re-described)	Observer did (or did not) like or accept the transgression (child uses words other than those used in the videos or by E); e.g., "Because she said, 'You broke Anya's doll; I don't think that's nice'"
	Moral character, evaluation, or norm	Observer is a good (or bad) person, observer's response to the transgression was good (or bad), or observer broke (or did not break) a moral norm; e.g., "Because she said the right thing" or "Because she is a better person"
	Enforcement (repeated) [analyses were conducted with this category scored as '1' and as '0']	Observer did (or did not) like or accept the transgression (child uses words that had been used in the videos or by E); e.g., "Because she said, 'Don't do that again'"
0	Own judgment	Child's own judgment of the transgression; e.g., "Because it's stupid when you break something"
	Own preference	Child's own preference for the observer; e.g., "Because I like her better"
	Object	Target object is damaged or observer noted that the object is damaged; e.g., "Because she said, 'You broke the doll'"
	Other, irrelevant, or uncodable	Response could not be put into any of the above categories (e.g., "I don't know"), was irrelevant (e.g., "Because that's more fun for me"), or was uncodable (e.g., the child's speech could not be understood)

sophisticated way and thus simply repeat the words they heard before? To account for both possibilities, we conducted two sets of analyses of children's justifications, one in which 'Enforcement (repeated)' was assigned a score of 1, and the other in which it was assigned a score of 0.

A score of 0 was assigned to all other justifications, including justifications that indicated that children had understood what had happened but that were not diagnostic, that is, they did not set one of the observers apart from the other. For instance, stating that the observer noted that the target object was broken received a score of 0 since the observer noted this in all scenarios. If a child did not provide a justification on a particular question, no score was assigned for that question. A second coder coded justifications of a random 25% of children from the transcriptions. Reliability was excellent, $\kappa = 0.90$.

3. Results

We first report results of the comprehension probes in order to provide information about how well children understood the content of the videos. We then report children's performance on test questions and the distribution of resources task.

Preliminary analyses revealed that for both age groups, there were no significant effects of gender; this factor was thus pooled for the main analyses. All reported *p*-values are two-tailed.

3.1. Comprehension probes

3.1.1. Comprehension probe 1

For all four videos, in response to the first comprehension probe, all children at both ages immediately responded that the observer did not think that the transgression was good. The only exception was one 4-year-old, who initially responded incorrectly in Phase 2 regarding the Non-enforcement video, but responded correctly after re-watching the video. Thus, all children grasped the basic premise of the videos and understood that both observers disapproved of the transgression.

3.1.2. Comprehension probe 2

Responses to the second comprehension probe indicated that most children at both ages understood what the observers said to the transgressor. Specifically, in Phase 1, when the observer enforced the norm, 20 of 24 4-year-olds (83.3%; 95% confidence interval (CI) [65.1%, 94.1%]) correctly stated right away that the observer said that the transgressor should never do that again (sign test, $p = 0.002$). Of the remaining 4 children, 2 responded correctly after re-watching the video. When the observer did not enforce the norm, 22 4-year-olds provided a response right away, of which 17 responded correctly that the observer said that the object is lying on the table (77.3%, CI [57.1%, 90.8%], sign test, $p = 0.017$). Five of the remaining 7 children responded correctly after re-watching the video. Results were very similar in Phase 2. In the Enforcement case, 23 children provided a response right away, of which 19 responded correctly (82.6%, CI [63.8%, 93.8%], sign test, $p = 0.003$). Two of the remaining 5 responded correctly after re-watching the video. In the Non-enforcement case, 23 children provided a response right away, of which 20 responded correctly (86.9%, CI [69.1%, 96.2%], sign test, $p < 0.0005$). Three of the remaining 4 responded correctly after re-watching the video.

The results of the 5-year-olds were very similar. In Phase 1, when the observer enforced the norm, all 24 5-year-olds correctly responded right away (100%, CI [90.2%, 100%], sign test, $p < 0.0005$). When the observer did not enforce the norm, 22 of 24 children (91.7%, CI [75.9%, 98.2%]) correctly responded right away (sign test, $p < 0.0005$), and the remaining 2 did so after re-watching the video. In Phase 2, all 24 5-year-olds (100%, CI [90.2%, 100%]) responded correctly right away in both the Enforcement and Non-enforcement cases (sign tests, both $ps < 0.0005$).

Altogether, in the Enforcement case, 17 of 24 4-year-olds and all 24 5-year-olds responded correctly right away in both phases, which was a significant difference (Fisher's Exact test due to small *ns* in some cells, $p = 0.009$, $\varphi = 0.41$). Similarly, in the Non-enforcement case, 15 of 24 4-year-olds and 21 of 24 5-year-olds responded correctly right away in both phases, which was also a significant difference (χ^2 [1, $N = 48$] = 4.00, $p = 0.046$, $\varphi = 0.29$). Thus, the 5-year-olds demonstrated a more robust and immediate grasp of the enforcer's and the non-enforcer's reactions to the transgressions than the 4-year-olds. Nonetheless, nearly all 4-year-olds responded correctly either right away or after re-watching the videos, and the few who did not received the critical information from the experimenter.

3.2. Test questions

3.2.1. Forced-choice questions

Preliminary analyses revealed that for both age groups, children's performance on the second comprehension probe (i.e., whether children responded correctly to the second comprehen-

sion probe or received the relevant information from the experimenter in each phase) was not significantly related to their performance on the test questions. This variable was thus not included in the analyses of test questions.

Preliminary analyses also indicated no significant effects of phase on children's responses to any of the four types of forced-choice test questions. We thus pooled children's responses to each type of question across the two phases, and assigned children a proportion score for each type of question, as follows: Children who responded in the hypothesized way to a particular question (e.g., 'Acted rightly') in both phases received a score of 1, children who responded in the hypothesized way to that question in only one of the two phases received a score of 0.50, and children who did not respond in the hypothesized way to that question in either phase received a score of 0. A given question was only included in analyses if children provided a response (indicating either the enforcer or non-enforcer) to that question. Thus, if for a particular question, children only provided a response in one phase, they received a 1 if that response was as hypothesized and a 0 if that response was not as hypothesized. [Note that assigning scores more conservatively, such that children also received a score of 0 for providing no response, resulted in a very similar pattern of results.]

To analyze children's responses to the forced-choice test questions, we first conducted one-sample Wilcoxon tests using the proportion score for each type of question as the dependent measure and test values of 0.50. These analyses revealed impressive results for the 5-year-olds: These older children drew the appropriate, hypothesized inferences on all four types of test questions, all $ps < 0.019$, r (effect size) values ranging from 0.34 to 0.52 (see Fig. 2). A one-sample t-test indicated that the proportion of the eight forced-choice test questions that 5-year-olds answered in the hypothesized way was significantly higher than 0.50, $M = 0.76$, $SD = 0.30$, $t(23) = 4.26$, $p < 0.0005$, *Cohen's* $d = 0.88$. Finally, a significant majority of 5-year-olds (18 of 24; 75%, CI [55.5%, 88.8%]) responded in the hypothesized way to more than half of the eight forced-choice test questions (sign test, $p = 0.023$).

The younger children, on the other hand, did not show a similarly high performance. First, one-sample Wilcoxon tests on the proportion scores for each type of question indicated that they did not draw any of the hypothesized inferences, all $ps > 0.616$ (see Fig. 2). A one-sample t-test indicated that the proportion of the eight forced-choice test questions that the 4-year-olds answered in the hypothesized way was also not significantly different from 0.50, $M = 0.52$, $SD = 0.32$, $t(23) = 0.37$, $p = 0.715$. Finally, only 11 of the 24 4-year-olds (45.8%, CI [27.3%, 65.3%]) responded in the hypothesized way to more than half of the forced-choice test questions (sign test, $p = 0.839$).

Comparing across ages using an independent-samples t-test, we found that the proportion of the eight forced-choice test questions answered in the hypothesized way was significantly higher among the 5-year-olds than the 4-year-olds, $t(46) = 2.61$, $p = 0.012$, *Cohen's* $d = 0.75$. Further, a chi-square test comparing the number of children who responded in the hypothesized way to more than half of the forced-choice test questions also revealed a significant difference, χ^2 [1, 48] = 4.27, $p = 0.039$, $\varphi = 0.30$.

3.2.2. Justifications

Justifications were only included in analyses if children had answered the preceding forced-choice test question in the hypothesized way (though note that including all justifications did not change the pattern of results). This resulted in 23 5-year-olds and 21 4-year-olds being included in the analyses, as one 5-year-old and three 4-year-olds did not answer any of the forced-choice test questions in the hypothesized ways.

Children's justifications were compared across the two age groups. When 'Enforcement (repeated)' was assigned a score of 1, nearly all 5-year-olds (20 of 23) were coded as providing at least one higher-level (score of 1) justification across all justification questions, indicating a sophisticated level of understanding and reasoning about the observers and their responses among these older children. A little more than half of the 4-year-olds (12 of 21) also provided at least one level-1 justification, but this proportion was significantly less than among the 5-year-olds, χ^2 [1, 44] = 4.92, $p = 0.027$, $\varphi = 0.33$. When 'Enforcement (repeated)' was assigned a score of 0, more than half of the 5-year-olds (13 of 23) still provided at least one level-1 justification, whereas less than half of the 4-year-olds (8 of 21) did so. However, this difference between age groups was no longer significant, $p = 0.222$.

3.3. Distribution of resources

The results of the distribution task mirrored those of the forced-choice test questions. Specifically, a one-sample t-test revealed that the 5.5-year-old children distributed significantly more than half (1.5) of the flowers to the enforcer, $M = 1.92$, $SD = 0.65$, $t(23) = 3.12$, $p = 0.005$, *Cohen's* $d = 0.64$. Moreover, 20 of the 24 5.5-year-olds (83.3%, CI [65.1%, 94.1%]) distributed 2 or 3 flowers to the enforcer, sign test, $p = 0.002$. On the other hand, a one-sample t-test revealed that the 4.5-year-old children did not distribute more than half of the flowers to the enforcer, $M = 1.46$, $SD = 0.66$, $t(23) = 0.31$, $p = 0.759$. Furthermore, only 11 of the 24 4.5-year-olds (45.8%, CI [27.3%, 65.3%]) distributed 2 or 3 flowers to the enforcer, sign test, $p = 0.839$.

Comparing the two age groups revealed significant differences. Specifically, an independent-samples t-test showed that the 5.5-year-olds distributed significantly more flowers to the enforcer than the 4.5-year-olds, $t(46) = 2.42$, $p = 0.020$, *Cohen's* $d = 0.70$. The number of 5.5-year-olds who distributed 2 or more flowers to the enforcer was also significantly greater than the number of 4.5-year-olds, χ^2 [1, 48] = 7.38, $p = 0.007$, $\varphi = 0.39$.

4. Discussion

In order for large-scale cooperation among unrelated individuals to be maintained, the norms of cooperation need to be enforced. However, norm enforcement is often a costly and risky endeavor, thus itself requiring a support mechanism. The study reported here provides evidence that this mechanism emerges fairly early in development. Specifically, by 5 years of age, children judge enforcers and enforcement more positively than non-enforcers and non-enforcement, personally prefer enforcers to non-enforcers, and provide more material rewards to enforcers than non-enforcers. Enforcers thus gain reputational and material benefits that prove advantageous in the long run, whereas non-enforcers do not gain such benefits. This provides a viable solution to the second-order problem of cooperation: Individuals who enforce cooperation norms on transgressors may pay costs and take risks to do so, but they also gain benefits from their group members in return (Barclay, 2006; Fessler & Haley, 2003; Gintis et al., 2001). These gains, along with the increased future cooperation of transgressors, may well benefit the enforcer sufficiently to encourage her to continue enforcing norms in the future, and may serve as an incentive for non-enforcers to begin enforcing norms (see Barclay, 2006; Nelissen, 2008). Our findings thus show that by the preschool years, children demonstrate the capacity (at least in a controlled lab setting) to contribute not only to first-order but also to second-order cooperation, and thus to maintain large-scale human cooperation in substantially more sophisticated ways than previously believed.

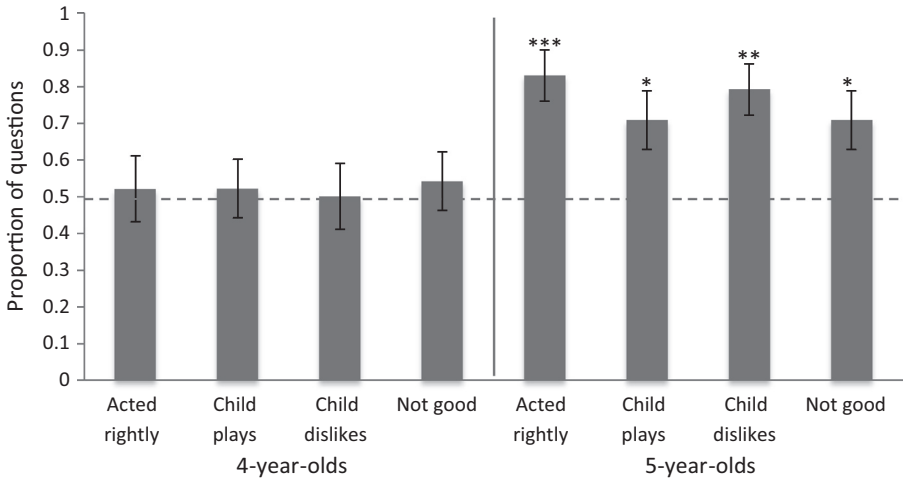

Fig. 2. Proportion of each type of forced-choice question answered in the predicted way across the two phases. The dashed line indicates the test value of 0.50. $^{*}p < 0.05$. $^{**}p < 0.005$. $^{***}p < 0.0005$.

Note that both the enforcers and non-enforcers expressed their disapproval of the transgressions. Children could thus be certain that the observers had similar and reasonable responses to the clear, unprompted transgressions featured in the videos.[1] The only substantial difference between enforcer and non-enforcer was thus in whether or not they enforced the norm on the transgressor. Note as well that the enforcer displayed displeasure during the enforcement, whereas the non-enforcer was neutral during the non-enforcement. This makes our results particularly striking, as 5-year-olds evaluated the enforcer positively despite the fact that she was superficially more unpleasant than the non-enforcer, suggesting that their judgments were not based on surface-level cues but on the content and import of the observers' behaviors. The fact that the majority of 5-year-olds justified their choices by appealing to reasons that were clearly relevant to the enforcement or non-enforcement (as evident in their justifications) also suggests that these older children attended to and grasped the significance of the observers' reactions and that their judgments were based at least in part on that information.

It should be noted that our scenarios featured norm enforcement in the form of reprimanding the transgressor. Although norm enforcement in the real world may often take this form, it can also take the form of active punishment (such as taking something away from or giving something unpleasant to a transgressor; see, e.g., Kenward & Östh, 2015). Since such punishment is more costly for the enforcer but also more harmful for the transgressor, 5-year-olds may find it more challenging to judge active punishment as positively as they judged the reprimanding in the present study. Indeed, even adults seem to disapprove of severe or overly aggressive punishment of norm violators (Eriksson et al., 2016). Thus, the capacity that 5-year-olds demonstrated in our study may not generalize to more severe forms of norm enforcement. Teasing these differences apart will be a fascinating avenue for future work.

Interestingly, the 4-year-olds in our study did not positively evaluate, prefer, or distribute more resources to the enforcer than non-enforcer. This is consistent with the finding from Kenward and Östh (2012) that 4-year-olds did not identify more with punishers than with non-punishers of first-order transgressors, and suggests that the propensity to value norm enforcement and enforcers may emerge between 4 and 5 years of age. This developmental shift is in line with other recent studies that have employed comparable methods and also found that whereas 5-year-olds demonstrate a robust grasp of rather sophisticated cooperation and group norms, 4-year-olds do not (Misch et al., 2014; Vaish et al., 2011). This developmental shift may reflect the fact that as children's interactions with strangers and peers and their experiences with group life (e.g., in kindergarten) increase, so too does their awareness of the importance of following and enforcing norms of cooperation (cf. Misch et al., 2014; Vaish et al., 2011).

Some alternative explanations for the 4-year-olds' performance need to be considered, however. First, perhaps 4-year-olds struggled with responding in the hypothesized ways to the test questions because they were unable to follow and keep track of the scenarios presented in the videos (both because they involved fairly long and complex social interactions and because each pair of videos that a given child watched featured two distinct target objects). Note, however, that the comprehension probes were designed to get around precisely this problem and to make sure that all children had the relevant information before they were asked the test questions. Indeed, nearly all 4-year-olds responded correctly to the comprehension probes, and the few who did not received the critical information from the experimenter. A second alternative is that the highly verbal nature of the task proved challenging for the 4-year-olds, who may have had difficulty understanding or providing verbal responses to the experimenter's questions. However, Vaish et al. (2011) tested this possibility and demonstrated that when asked very similar test questions about simpler scenarios (i.e., scenarios involving less sophisticated cooperation norms), 4-year-olds did respond in the hypothesized ways, and their performance was comparable to that of 5-year-olds. Moreover, note that the 4-year-olds in our study also did not perform as hypothesized on the distribution task, which was a non-verbal and thus arguably easier task. It thus seems unlikely that the 4-year-olds in our study struggled with the verbal nature of

[1] We would like to note here that prior to this study, we conducted a study using a very similar procedure with the exception that in the videos, neither observer privately expressed her disapproval of the transgression; rather, the enforcer expressed her disapproval of the transgression while enforcing the norm, whereas the non-enforcer did neither. The performance of the 5-year-olds in that study was similar to that of the 5-year-olds in this study. In addition, the 4-year-olds in that study also partially performed at statistically significant levels. However, reviewers drew our attention to the potential confound that children may have relied on the enforcer's disapproval of the transgression rather than the enforcement. We thus conducted the present study, with new videos in which both observers privately express their disapproval (so as to feature disapproval from both observers but without the disapproval being misperceived as enforcement of the norm upon the transgressor).

the task; rather, it seems likelier that as a group, 4-year-olds may not yet value and prefer enforcers.

It is noteworthy, however, that several 4-year-olds justified their choices on the forced-choice questions by appealing to enforcement-related reasons. This suggests that at least some children of this age are beginning to attend to and grasp the importance of norm enforcement. It will be interesting for future work to explore why some children demonstrate this understanding earlier than others and what social-cognitive and socialization factors this depends upon. For instance, children's ability to engage in higher-order reasoning, parents' focus on norm following and enforcement, as well as children's exposure to norm-heavy environments such as preschool likely play an important role in the emergence of this understanding.

Relatedly, there is likely to be some cross-cultural variation in children's responses to norm enforcement, both in age of emergence and the nature of the responses. Although no cross-cultural work exists on this topic yet, we do know that there is cross-cultural variation in both adults' and children's punishment of first-order transgressors that depends on the norms and institutions of the groups (e.g., Henrich et al., 2006; Robbins & Rochat, 2011). It seems plausible that this variation carries over into adults' and children's responses to second-order cooperators and non-cooperators as well – though one may predict that because every human group needs to sustain cooperation, every group should show some degree or form of valuing second-order cooperators. This is a fascinating direction for future research.

A further fascinating question concerns the mechanism underlying children's judgments, that is, whether children were responding positively to the enforcer, negatively to the non-enforcer, or both. Existing work with adults is inconclusive on this point, with some work indicating that rewards sustain second-order cooperation and other work suggesting that punishment does (see, e.g., Boyd & Richerson, 1992; Henrich & Boyd, 2001; Kiyonari & Barclay, 2008; Nelissen, 2008). Our study design cannot tease these possibilities apart as we compared children's responses to enforcers versus non-enforcers. Future work will need to compare children's responses to enforcers versus neutral (e.g., naïve) individuals as well as non-enforcers versus neutral individuals in order to provide insights on this question (see, e.g., Misch et al., 2014; Vaish, Grossmann, & Woodward, 2008).

Our findings effectively show that preschool-aged children have the capacity to engage in third-order cooperation. But this raises a further question: How is third-order cooperation sustained? More generally, it raises the infinite regress problem, namely, that if norm enforcement and punishment were themselves supported by rewards and/or punishment, then those rewards and punishment would require a higher order of rewards and punishment in order to be maintained, and so on *ad infinitum* (Henrich & Boyd, 2001; Sripada, 2005). Some authors have proposed that beyond the second or third order of reward and punishment, additional orders of reward and punishment are likely maintained not through still greater orders of reward and punishment but rather by the reputational benefits gained by the enforcer. That is, unlike costly rewards and punishment, trusting and respecting the enforcer do not require further levels of explanation, since it is in everyone's best interest to value norm enforcers as trustworthy partners and avoid cheating them in order to avoid sanctions (Barclay, 2006). Interestingly, it has been argued that due to cognitive limitations, rewards and punishment likely do not go beyond the second or at the most the third order anyway (e.g., Sripada, 2005). Although our study cannot address these issues, this is an interesting and important avenue for future work.

Needless to say, second-order cooperation is not sustained solely through rewards and punishment. One additional mechanism may be conformity (e.g., Boyd & Richerson, 1985; Henrich & Boyd, 2001), whereby if a few enforcers gain positive reputations and status, and are thus more successful group members than non-enforcers, then bystanders tend to copy the enforcers rather than the non-enforcers; with time, enforcement is established as the dominant behavior and others naturally tend to conform to this norm, making non-enforcement the exception (see Salali et al., 2015, for relevant work). Moreover, as enforcement becomes widespread in a group, the number of norm violators is rapidly reduced such that enforcement rarely needs to occur at all; this dramatically lowers the costs of enforcement, making it more likely to stabilize (Sripada, 2005). These and other mechanisms likely work along with reputation to stabilize norm enforcement and punishment.

It is also important to note that the mechanisms that sustain norm enforcement need not be the same as the enforcers' motivations. That is, the benefits that enforcers gain may not be the reason that they enforce norms; rather, they may well enforce norms out of genuine moral indignation or concern for others' (or the group's) welfare (Sripada, 2005). Equally, bystanders may not judge enforcers positively in order to ensure that cooperation succeeds in their group but because it is in their selfish interest to recognize and value enforcers as more reliable and trustworthy cooperation partners than non-enforcers (Barclay, 2006). These proximate motivations notwithstanding, as long as bystanders value enforcers and as long as enforcers benefit from these positive evaluations, second-order cooperation can succeed.

In conclusion, we have shown here that from a young age, people value, prefer, and provide more resources to those who enforce norms on others more than those who do not enforce norms. The costly and risky norm enforcement that allows large-scale cooperation to succeed is likely sustained, at least in part, through these reputational and material benefits that enforcers gain and non-enforcers lose (Barclay, 2006; Fessler & Haley, 2003; Nelissen, 2008). Humans thus become well equipped from early on to uphold complex and large-scale cooperation.

Acknowledgements

This research was partially supported by a grant from the DFG Excellence Cluster 302 (Languages of Emotions) in cooperation with the Free University Berlin. Amrisha Vaish was supported by a Dilthey Fellowship provided by the Volkswagen and the Fritz Thyssen Foundations. We thank the daycare centers and children for their friendly cooperation, Stefanie Gutknecht, Liane Jorschik, Kathleen Scholz, Gesa Volland, Katja Weber, and the MPI-EVA Multimedia staff for help with creating the videos, Kristina Schilke for help with coding, and the MPI-EVA IT staff for technical support.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.cognition.2016.04.011.

References

Barclay, P. (2006). Reputational benefits for altruistic punishment. *Evolution and Human Behavior, 27*, 325–344.

Boyd, R., & Richerson, P. J. (1985). *Culture and the evolutionary process*.Chicago: The University of Chicago Press.

Boyd, R., & Richerson, P. J. (1992). Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology, 13*, 171–195.

Eriksson, K., Andersson, P. A., & Strimling, P. (2016). Moderators of the disapproval of peer punishment. *Group Processes & Intergroup Relations, 19*, 152–168. http://dx.doi.org/10.1177/1368430215583519.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature, 415*, 137–140.

Fessler, D. M. T., & Haley, K. (2003). The strategy of affect: Emotions in human cooperation. In P. Hammerstein (Ed.), *Genetic and cultural evolution of cooperation* (pp. 7–36). Cambridge, MA: MIT Press.
Frank, R. H. (1988). *Passions with reason: The strategic role of emotions*.New York: W. W. Norton.
Gintis, H., Smith, E. A., & Bowles, S. (2001). Cooperation and costly signaling. *Journal of Theoretical Biology, 213*, 103–119.
Hamlin, J. K., Wynn, K., Bloom, P., & Mahajan, N. (2011). How infants and toddlers react to antisocial others. *Proceedings of the National Academy of Sciences, 108*, 19931–19936.
Henrich, J. P. (2004). *Foundations of human sociality: Economic experiments and ethnographic evidence from fifteen small-scale societies*.Oxford: Oxford University Press.
Henrich, J., & Boyd, R. (2001). Why people punish defectors: Weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *Journal of Theoretical Biology, 208*(1), 79–89.
Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., et al. (2006). Costly punishment across human societies. *Science, 23*, 1767–1770.
Horita, Y. (2010). Punishers may be chosen as providers but not as recipients. *Letters on Evolutionary Behavioral Science, 1*(1), 6–9.
Kenward, B., & Dahl, M. (2011). Preschoolers distribute scarce resources according to the moral valence of recipients' previous actions. *Developmental Psychology, 47*(4), 1054–1064.
Kenward, B., & Östh, T. (2012). Enactment of third-party punishment by four-year-olds. *Frontiers in Psychology, 3*. http://dx.doi.org/10.3389/fpsyg.2012.00373.
Kenward, B., & Östh, T. (2015). Five-year-olds punish antisocial adults. *Aggressive Behavior, 41*, 413–420. http://dx.doi.org/10.1002/ab.21568.
Kiyonari, T., & Barclay, P. (2008). Cooperation in social dilemmas: Free riding may be thwarted by second-order reward rather than punishment. *Journal of Personality and Social Psychology, 95*(4), 826–842.
Milinski, M., Semmann, D., & Krambeck, H. (2002). Reputation helps solve the "tragedy of the commons". *Nature, 415*, 424–426.
Misch, A., Over, H., & Carpenter, M. (2014). Stick with your group: Young children's attitudes about group loyalty. *Journal of Experimental Child Psychology, 126*, 19–36.
Nelissen, R. M. A. (2008). The price you pay: Cost-dependent reputation effects of altruistic punishment. *Evolution and Human Behavior, 29*, 242–248.
Nowak, M. A. (2006). Five rules for the evolution of cooperation. *Science, 314*, 1560–1563.
Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature, 437*, 1291–1298.
Riedl, K., Jensen, K., Call, J., & Tomasello, M. (2015). Restorative justice in children. *Current Biology, 25*, 1731–1735. http://dx.doi.org/10.1016/j.cub.2015.05.014.
Robbins, E., & Rochat, P. (2011). Emerging signs of strong reciprocity in human ontogeny. *Frontiers in Psychology, 2*. http://dx.doi.org/10.3389/fpsyg.2011.00353.
Salali, G. D., Juda, M., & Henrich, J. (2015). Transmission and development of costly punishment in children. *Evolution and Human Behavior, 36*, 86–94. http://dx.doi.org/10.1016/j.evolhumbehav.2014.09.004.
Smetana, J. G., Schlagman, N., & Adams, P. W. (1993). Preschool children's judgments about hypothetical and actual transgressions. *Child Development, 64*, 202–214.
Sober, E., & Wilson, D. S. (1998). *Unto others: The evolution and psychology of unselfish behavior*.Cambridge, MA: Harvard University Press.
Sripada, C. S. (2005). Punishment and the strategic structure of moral systems. *Biology and Philosophy, 20*, 767–789.
Vaish, A., Carpenter, M., & Tomasello, M. (2010). Young children selectively avoid helping people with harmful intentions. *Child Development, 81*(6), 1661–1669.
Vaish, A., Carpenter, M., & Tomasello, M. (2011). Young children's responses to guilt displays. *Developmental Psychology, 47*(5), 1248–1262.
Vaish, A., Grossmann, T., & Woodward, A. (2008). Not all emotions are created equal: The negativity bias in social-emotional development. *Psychological Bulletin, 134*(3), 383–403.
Vaish, A., Missana, M., & Tomasello, M. (2011). Three-year-old children intervene in third-party moral transgressions. *British Journal of Developmental Psychology, 29*, 124–130.
Wedekind, C., & Milinski, M. (2000). Cooperation through image scoring in humans. *Science, 288*, 850–852.
Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology, 51*(1), 110–116.