MORAL DUTY AND EQUALISATION CONCERNS MOTIVATE CHILDREN'S THIRD-PARTY PUNISHMENT

Although children enact third-party punishment, at least in response to harm and fairness violations, much remains unknown about this behaviour. We investigated the tendency to make the punishment fit the crime in terms of moral domain; developmental patterns across moral domains; the effects of audience and descriptive norm violations; and enjoyment of inflicting punishment. We tested 5- to 11-year-olds in the UK (N = 152 across two experiments, 55 girls and 97 boys, predominantly white and middle-class). Children acted as referees in a computer game featuring teams of players: as these players violated fairness or loyalty norms, children were offered the opportunity to punish them. We measured the type (fining or banning) and severity of punishment children chose and their enjoyment in doing so. Children only partially made the punishment fit the crime: they showed no systematic punishment choice preference for disloyal players, but tended to fine rather than ban players allocating resources unfairly – a result best explained by equalisation concerns. Children's punishment severity was not affected by audience presence or perpetrators' descriptive norm violations, but was negatively predicted by age (unless punishment could be used as an equalisation tool). Most children did not enjoy punishing, and those who believed they allocated real punishment reported no enjoyment more often than children who believed they pretended to punish. Contrary to predictions, retribution was not a plausible motive for the observed punishment behaviour. Children are likely to have punished for deterrence reasons or because they felt they ought to.

Keywords: third-party punishment; children; affective states; audience effects; descriptive and injunction norm violations; moral domains

Punishment is a behaviour intended to impose costs upon transgressors of norm violations, and can come in a wide range of forms: from verbal and physical confrontations to social exclusion and subtraction of economic resources (Molho, Tybur, Van Lange, & Balliet, 2020). Consequent costs for the punisher may include decrease in social support, psychological wellbeing and/or material resources (Adams & Mullen, 2012; van den Berg, Molleman, & Weissing, 2012), or be essentially absent, in the case of anonymous acts (Klempka & Stimson, 2014). Furthermore, punishment can be classified depending on whether it targets self- or other-relevant transgressions: in second-party punishment (2PP) the wrongdoer is punished by the victim of the norm violation, while in third-party punishment (3PP) the wrongdoer is punished by a bystander to the norm violation. Whereas the former process is present in other animal species, the latter seems to be uniquely human (Riedl, Jensen, Call, & Tomasello, 2012). Unlike second-party punishers, third-party punishers may suffer a cost apparently to the benefit of others (Jensen, 2010). This opens fascinating and unresolved questions as to how processes of biological or cultural selection could have favoured the evolution of 3PP (Chudek & Henrich 2011; Wilson & Sober, 1994), and even discussions as to whether costly 3PP is even a common phenomenon (Guala, 2012; Balafoutas, Nikiforakis, & Rockenbach, 2014).

This work, however, focuses on the proximate mechanisms of 3PP across development rather than on its adaptive functions (Tinbergen, 1963). In common with much of the developmental literature reviewed below, we do not assume that 3PP is by definition costly to the punisher. Rather, we are interested in the psychological mechanisms involved when children decide to enact a cost to an individual who has transgressed against a third party, in part independently of the issue of cost to the child. We now discuss psychological mechanisms that have been identified to be important in adults – retribution, deterrence, reputation and equalisation concerns – before outlining what is known about children’s 3PP.

Adults assign 3PP to transgressors even in scenarios where there is no chance for the group to benefit from a potential change in the targets’ behaviour (Crockett, Özdemir, & Fehr, 2014). Not only do people enact 3PP in one-shot interactions (Fehr & Gächter, 2000, 2002), but during repeated-interaction experiments they even show higher levels of 3PP in the last rather than first rounds (Gächter, Renner, & Sefton, 2008, as cited by Raihani & Bshary, 2019). This suggests that people are motivated by *retribution*, i.e. 3PP for the sake of giving wrongdoers their “just deserts”, without any further instrumental reason.

Other accounts argue that 3PP has a *deterrent motivation* to prevent misbehaviours from occurring to oneself (Delton & Krasnow, 2017; Krasnow, Delton, Cosmides, & Tooby, 2016) or to people the punisher has a welfare stake in, such as kin, friends or in-group members (Ericksen, & Horton, 1992; Lieberman & Linke, 2007). 3PP could thus be viewed as a bargaining chip in social exchanges: individuals indeed avoid making punitive efforts to reform uncooperative behaviour targeting exclusively unknown others (Krasnow, Cosmides, Pedersen, & Tooby, 2012).

Relative payoff concerns can also offer an explanation for third-party punishers’ sensitivity to inequality. Indeed, people who engage in the costly reduction of payoff differences between group members, when inequalities are the product of chance, are likely to be the same people who enact 3PP against individuals unwilling to cooperate in the group (Johnson, Dawes, Fowler, McElreath, & Smirnov, 2009). Furthermore, 3PP of unfairness seems to be motivated more by envy of the wrongdoer’s higher payoff than by moralistic anger at the experience of the victim of unfairness (Pedersen, Kurzban, & McCullough, 2013).

Third-party punishers can also accrue social benefits from their intervention via *reputational gains*. There is indication that punishment on behalf of strangers is practised to escape bystanders’ negative judgements (Pedersen, McAuliffe, & McCullough, 2018). Individuals invest more resources in enacting 3PP when they are aware their decisions will be communicated to an audience than when their decisions will remain anonymous (Kurzban, DeScioli, & O'Brien, 2007). 3PP might function as a mechanism to signal punishers' cooperative qualities, such as trustworthiness (Jordan, Hoffman, Bloom, & Rand, 2016), concern about group's shared values and social standing of the victim (Okimoto & Wenzel, 2011), as well as commitment to impartiality and fairness (Baumard, André, & Sperber, 2013; Nelissen, 2008). Additionally, 3PP could also work as a costly signal of formidability to dissuade observers from implementing any exploitive intentions they might have (Raihani & Bshary, 2015). Thus, 3PP might be akin to a strategy to assert dominance (Sylwester, Hermann, & Bryson, 2013).

Third-party punishment in childhood

Although behavioural research into 3PP involving adults is well-established, less is known about such punitive behaviour in children. An appetite for bad things to happen to bad individuals is present from very early on: 8-month-old infants prefer third parties who punish (instead of helping) antisocial individuals; 19-month-old toddlers prefer to personally enact 3PP over help towards antisocial individuals (Hamlin, Wynn, Bloom, & Mahajan 2011). A desire to punish wrongdoers is evident even when children are not explicitly encouraged to punish (Kenward & Östh, 2012) or when imposition of a cost upon transgressors is not framed as punishment (Kenward & Östh, 2015). Some children engage in 3PP even when they have to pay a social cost (Kenward & Östh, 2015) or an economic cost (Gummerum & Chu, 2014; McAuliffe, Jordan, & Warneken, 2015; Robbins & Rochat, 2011; Salali, Juda, & Henrich, 2015). Children intervene as third-party punishers when they observe a range of norm violations involving issues of fairness (Gummerum & Chu, 2014; Gummerum, López-Pérez, Van Dijk, & Van Dillen, 2019; Jordan, McAuliffe, &

Warneken, 2014; McAuliffe et al., 2015; Robbins & Rochat, 2011; Salali et al., 2015; Smith & Warneken, 2016) or harm (Hamlin et al., 2011; Kenward & Östh, 2012, 2015; Van de Vondervoort & Hamlin, 2018). Types of punishment investigated have mainly consisted of children withholding or taking away resources from transgressors (Gummerum & Chu, 2014; Gummerum et al., 2019; Hamlin et al., 2011; Jordan et al., 2014; McAuliffe et al., 2015; Riedl, Jensen, Call, & Tomasello, 2015; Robbins & Rochat, 2011; Salali et al., 2015), or inflicting them harm (Kenward & Östh, 2015; Marshall, Gollwitzer, Wynn, & Bloom, 2019). It has been demonstrated that 3PP rates in children increase in response to modelling (Salali et al., 2015) and with age (Jordan et al., 2014; McAuliffe et al., 2015; Salali et al., 2015), but that 3PP severity decreases with age (Gummerum, Takezawa & Keller, 2009). There is also indication that gender (Kenward & Östh, 2015), culture (Robbins & Rochat, 2011) as well as authority and ingroup-outgroup dynamics influence punitive behaviour (Gummerum et al., 2009; Jordan et al., 2014; Yudkin, Van Bavel, & Rhodes, 2019). Moreover, pre-schoolers prefer victim restoration over 3PP of transgressors (Riedl et al., 2015). There is also some indication that children's explanations of the reason to intervene as third-party punishers incorporate deterrent and pedagogical elements (Yudkin et al., 2019). Finally, the experience of negative emotions does not appear to motivate 3PP decisions in children (Gummerum et al., 2019).

Current study

In summary, although it has been shown that children do engage in 3PP in experimental contexts, because of the relative recency of this field, most studies have focussed on establishing this simple fact and examining relatively straightforward predictors of 3PP such as age, cost and modelling effects. As such, much remains to be known about the proximate mechanisms that regulate children's 3PP reactions in these contexts. This paper will present two experiments that were designed to investigate the following relevant issues: whether children tend to fit the kind of punishment to the kind of moral violation in terms of moral domain (Experiments 1-2); whether they punish violations of descriptive norms (what is commonly done) as well as violations of moral norms (Eriksson, Strimling, & Coultas, 2015) (Experiment 1); whether their 3PP responses to different types of moral violations are affected by age (Experiments 1-2) and the presence of an audience (Experiment 2); and what affective states they experience in enacting 3PP (Experiment 2). In order to fill these gaps in knowledge, a two-player cooperative spaceship computer game – called *MegaAttack* – was developed to be used in experiments with primary school-aged children (ages 5–11 years). In *MegaAttack* players belonging to the same team cooperate with one another against computer-controlled enemies. After having had a chance at playing cooperatively in a team with the experimenter in a face-to-face interaction (offline playing phase) as game familiarisation, children changed role from players to referees whose job was to judge supposed internet players' behaviour during the game (online refereeing phase). Children policed misbehaviours as unaffected bystanders, on behalf of the victims, but they were never victims themselves. Children did not have to pay any economic or social costs to engage in 3PP.

Studies assessing the ecological validity of experimental games employed with adults show contrasting results: while some studies have found correlational evidence between behaviours in experimental settings and behaviours in real-world situations (e.g., Benz & Meier, 2008; Gervais, 2017), others have not (e.g., Galizzi & Navarro-Martínez, 2018; Winking & Mizer, 2013). However, our intent was not to devise an experimental game fully generalisable to contexts outside the laboratory, but to test hypotheses about children's punitive preferences (Guala, 2012; Pisor, Gervais, Purzycki, & Ross, 2019). We specifically wanted to produce causal knowledge about the cognitive and affective processes moderating 3PP, but for causal relations to be isolated we needed controlled conditions that are achievable only in experimental games (Falk & Heckman, 2009). These methods are not without their limitations. For example, to be able to explore 3PP we framed our game and defined the set of behavioural choices available to the children in such a way to maximise the chances that they would respond to norm violations with 3PP (for example by not requiring children to pay a cost to punish, see Pedersen et al., 2018). However, most of our hypotheses do not relate to whether children would punish, but rather to details of how they punish. While we are thus cautious of not conflating (experimental) perceived expectations with (real-life) internal motivations as drivers of behaviour (List, 2007; Levitt & List, 2007), we also argue that moderators of elicited punishment behaviour might also be relevant for considering spontaneous punishment behaviour (similarly to how an experiment on lying can be revealing of mechanisms of lying even though participants are asked to lie; Vrij, Granhag, Mann, & Leak, 2011).

Experiment 1

Social norm classifications

An important debate about moral norms concerns the contraposition between monism and pluralism, where the former considers all moral concerns as manifestations of a unique moral domain (Baumard et al., 2013; Schein & Gray, 2018), while the latter asserts that there is more than one moral domain. Early pluralist theories (e.g., Shweder, Much, Mahapatra & Park, 1997) have been built on by theories such as “Moral Foundations Theory”. Moral Foundations Theory includes five moral foundations: c*are/harm* and *fairness/cheating* (individualising foundations); *loyalty/betrayal*, *authority/subversion* and *sanctity/degradation* (binding foundations) (Graham et al., 2013). Graham and colleagues (2013) have pointed out that research in developmental moral psychology has hardly begun when it comes to domains other than harm and fairness.

In the context of pluralistic theories the nature of the link between transgressions relating to different moral domains and consequent punitive motivations has not been clarified. We propose two rival hypotheses: general vs specific punishment behaviour motivations. According to the *specific motivation hypothesis*, transgressions of different domains lead to different types of punishment motivation, potentially motivating different types of punishment behaviour (the "punishment fits the crime" hypothesis, Figure 1A). According to the *general motivation hypothesis*, instead, detection of transgressions in different domains leads to a generic sense that a transgression has occurred and thus different types of transgression activate the same type of punishment motivation (Figure 1B).

Given the absence of literature on children's punitive attitudes towards violations apart from those related to harm and fairness, and the lack of literature comparing children's punishment of violations in different domains, we investigated whether children tend to react differently to different types of moral norm violations. We thus investigated for the first time children's punitive responses to violations of what Moral Foundations Theory considers a binding foundation – loyalty. In order to put the specific motivation hypothesis to the test, we predicted that unfairness in resource distribution might be more likely to motivate economic punishment, whereas disloyalty might be more likely to motivate social punishment such as ostracism. We also predicted that this tendency to match the type of punishment with the type of moral violation would vary with age because of potential developmental tendencies to cognitive differentiation or integration (Siegler & Chen, 2008).

Another norm classification approach – proposed by both Cialdini, Reno & Kallgren (1990) and Bicchieri (2005) – distinguishes between *descriptive norms* (i.e., what people typically do) and *injunctive norms* (i.e., what people think that ought to be done). Based on recent evidence that children negatively evaluate descriptive norm violations (Roberts, Guo, Ho, & Gelman, 2018), one might expect them to elicit also punitive sentiments. We thus investigated whether descriptive norm violations would increase the severity of 3PP allocated for moral norm violations. Results of this investigation were somewhat inconclusive and further introduction and discussion of the issue is therefore provided in Supplementary Information (section S4). Because substantial variance in punishment severity is typically explained by judgements of transgression severity (Alter, Kernochan, & Darley, 2007), we measured and controlled for transgression severity judgements when modelling punishment severity.

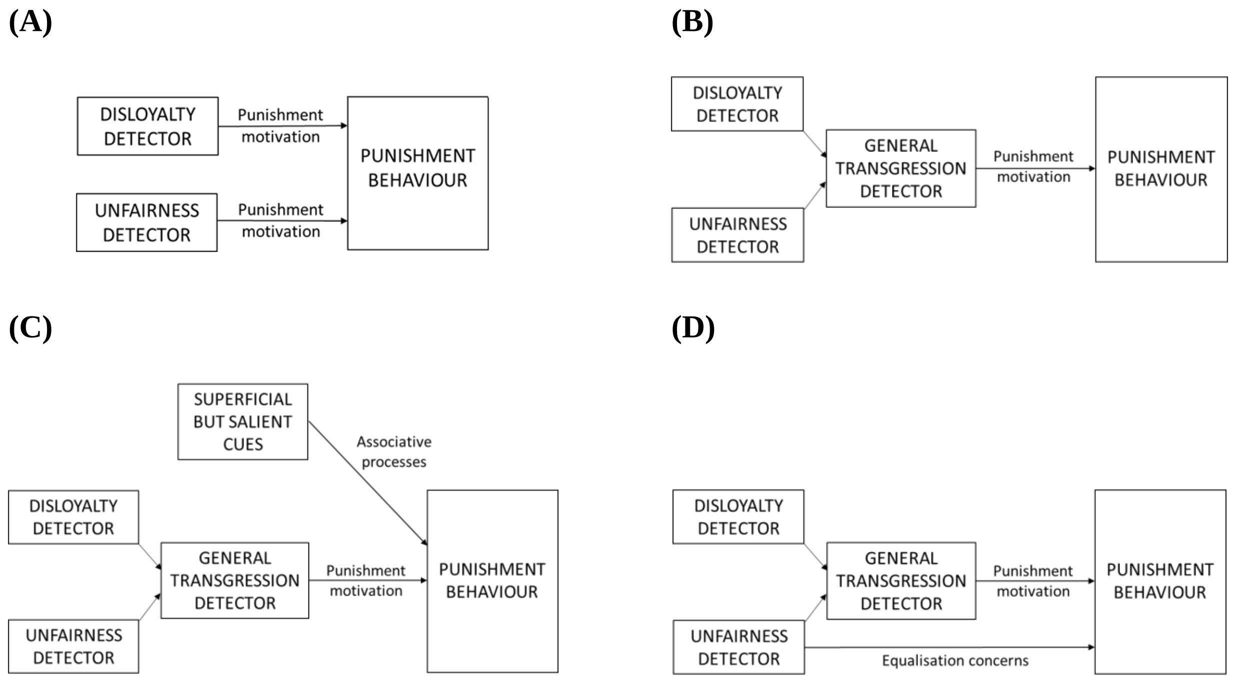

***Figure 1*. Hypothesised punishment motivations illustrating the relationship between transgressions in different moral domains and consequent punitive outcomes.** A) Specific motivation hypothesis. B) General motivation hypothesis. C) Associative hypothesis. D) General motivation plus equalisation hypothesis.

Age effect on third-party punishment

In the developmental literature the probability of children engaging in 3PP has been shown to increase with age, across different countries and types of moral scenarios. Specifically, this upward developmental pattern in 3PP rates has been detected in children who watched unfair allocations made during a Triadic Dictator Game. This economic paradigm has been adopted by Jordan, McAuliffe & Warneken (2014) with US children (age groups: 6 and 8 years of age); by Salali, Juda & Henrich (2015) with Canadian children (age range: 3 to 8 years of age); and by McAuliffe, Jordan & Warneken (2015) with US children (age groups: 5 and 6 years of age). Similarly, Smith & Warneken (2016) demonstrate an increasing tendency in US children between 4 and 10 years of age to use resource distributions to disadvantage transgressors. By contrast, the Triadic Dictator Game study conducted by Gummerum et al. (2009) revealed a downward developmental pattern in punitiveness. Their participants were recruited in Germany, and were both children (age groups: 7 and 11 years of age) and adults (mostly university students). Children proved to be more punitive third-parties than adults. Notably, in this case punitiveness was operationalised as 3PP severity rather than 3PP rates.

However, since the majority of the literature about the development of punitiveness indicated an upward pattern, we predicted we would detect the same in Experiment 1 even though we measured children's punitiveness in terms of 3PP severity instead of 3PP rates. Furthermore, previous studies have never analysed how punitiveness develops across different moral domains, as they were focused on issues of either unfairness or harm, but never on both at the same time. Therefore, in order to test the generalisability of those findings, we explored whether the development of 3PP severity would be affected by the moral domain of the transgressions (disloyalty vs unfairness) children witnessed.

Method

Materials. The *MegaAttack* game was programmed in LÖVE, an open-source game development environment utilising the LUA programming language, and run on a laptop computer which was taken to test locations. Headphones were used so that the audio could be clearly heard in noisy environments like science fairs. In the test trials, participants saw recordings of games that they were told were being played live by internet players. The descriptive norm violation was operationalised as a protective-shield colour-choice made in contrast with what was preferred by all other player-avatars displayed in the game. The loyalty violation was operationalised as a refusal to protect a team member who was under deadly attack. The fairness violation was operationalised as an unfair distribution of game resources (gems).

Sample. Participants were 72 primary school-aged children (*mean age*: 8.83 years; *SD* = 1.81 years; *age range*: from 5.45 years to 11.95 years; 32 females and 40 males) tested in a diverse range of settings – one museum, one primary school and two science fairs – but the whole testing phase took place in the same medium-sized English city (from June to October 2017). Power analyses were not performed because of the lack of previous data on which to base effect size expectations, so we allowed logistical constraints to determine effect sizes. The study was approved by the Oxford Brookes University Research Ethics Committee (Study Number 171101, Children's social judgement in a computer game).

Thirty-five of 72 parents (18 fathers; 15 mothers; 2 unspecified) partially or fully completed a socio-demographic questionnaire, indicating that Experiment 1's sample came predominantly from a middle-class background (the median yearly family income was £60,000; one out of the 35 respondents preferred not to declare) with a high education level (88.57% of the respondents had at least a Bachelor's degree), and was heterogeneous in terms of nationality (parents' nationality: 23 British, 10 non-British, 2 unspecified). Data on racial identity was not systematically collected, but the sample was predominantly white.

Design. We adopted a 2x2 fully within-subject design in which the factors were *descriptivity* (descriptive norm conformity; descriptive norm violation) and *type of moral transgression* (fairness transgression; loyalty transgression), see Table 1. We ran one trial in each condition combination, with each trial featuring two unique players, one violator and one non-violator. In the resulting four trials a moral transgression always occurred (either a fairness or loyalty norm violation), and a descriptive norm violation either did or did not occur, with these variables counterbalanced. Two irrelevant variables were counterbalanced across participants: the descriptively normative colour choice (red or blue), and the order of trials. Order with respect to descriptive norm violation/conformity was AABB or BBAA, and with respect to loyalty/fairness transgression was ABAB or BABA, counterbalanced (four possible order variants, see Supplementary Information – Table S1 for details). Each test-trial featured a different pair of player avatars (different animals inside space-ships).

The dependent variables measured were: *judgement of transgression severity* (5 ordinal levels: from "just a little bad" to "super bad", Figure S1 in Supplementary Information); *type of punishment* (2 categorical levels: economic, loss of gems as an in-game resource vs social, banning from the game, Figure S2 in Supplementary Information); *severity of punishment* (6 ordinal levels for both social punishment and economic punishment, ranging from no punishment to 1 day of ban or a 100 gem fine, Figure S2 in Supplementary Information).

***Table 1.* List of key independent variables for each experiment with details of the levels for each variable, plus indication of whether the variables were manipulated within- or between-subjects.**

Independent variable	Experiment	Variable's levels	Manipulation
Descriptivity	1	Descriptive norm violation; descriptive norm conformity	Within-subjects
Type of moral transgression	1 - 2	Fairness norm transgression; loyalty norm transgression	Within-subjects
Audience	2	Present; absent	Within-subjects
Punishment opportunity	2	Real; warning; pretend	Between-subjects

Procedure. The procedure was divided into three phases (see full script in Supplementary Information – section S1 for further details): (1) **Familiarisation**, further subdivided into an **offline playing familiarisation** and a purportedly **online refereeing familiarisation**; (2) Four purportedly **online test trials**; (3) **Final questions**. Familiarisation and Final questions were identical for all participants.

Parents of all children gave informed written consent for them to take part in the experiment. Children were tested by a single experimenter, seated at a laptop, with any accompanying adults engaged in other activities (for example filling in the questionnaire). The procedure began with the experimenter explaining to the children that the experiment consisted of playing offline and refereeing online a newly devised computer game called *MegaAttack*.

The **playing familiarisation** was organised into four short game bouts, aimed at establishing for the participant that standard moral norms applied to the game, with respect to issues of team loyalty and fairness in resource distribution. At the beginning, the child and the experimenter were automatically assigned shields of the same colour (the one that in test trials would be descriptively normative). They then flew space-ships, playing together as a team, defending themselves by shooting robot attackers, and collecting gems that initially went into a communal store but were manually divided between the players by one of the players at the end of the game bouts.

Each of the **four bouts of the playing familiarisation** was constituted by a gem collection stage (45 seconds) followed – from the second bout onwards – by a gem division stage (15 seconds). The first bout had no gem division, for ease of introducing the game; the child decided how to split the gems at the end of the second bout, and the experimenter split the gems at the end of the third and fourth bouts. Both times, the experimenter split the gems equally between herself and the child, thus demonstrating that fair division was normal. A team-loyalty norm was demonstrated when the experimenter came to the aid of the child when the child's space-ship was in danger of being destroyed during a mega-attack, a sudden event in which an overwhelming number of enemies surrounded and attacked the child's space-ship at the same time (during the fourth bout). After the playing familiarisation bouts, the participant was told they were to **referee the game** by judging the behaviour of some internet players (the two players represented on the screen were described as having connected to the game live via the internet, but the games displayed were actually pre-recorded).

Differently from the bouts in the playing familiarisation, in each bout the child had to referee (one refereeing familiarisation bout and four test trial bouts) a shield-choice stage (5 seconds) preceded the gem collection and division stages, in which each player chose either a red or blue shield. At the beginning of the **refereeing familiarisation** bout the descriptive norm was introduced to the child: the experimenter explicitly said that internet players commonly chose a specific shield colour over another one (red or blue counterbalanced across participants). To support this claim, the child was invited to pay attention to the shield colour used by 28 additional avatars outside the game arena, on the edge of the screen, presented as internet players that were waiting to play. In the refereeing familiarisation bout no norms were violated by the two players: both players chose the common over the uncommon shield colour and both players were loyal and fair to each other. For this reason the child was expected to conclude that no misbehaviours had occurred.

The refereeing familiarisation was followed by **four test trials** (each one game bout) in which the child saw a combination of descriptive and moral norm-violations (as outlined above in the section dedicated to the experimental design) and heard the narration of such actions from a live-streamer (commentator) presented as live but actually pre-recorded (note that live internet-game commentary is now a common phenomenon that many children are familiar with; Sjöblom & Hamari, 2017). Two different male voice-overs were used, counterbalanced across participants. Children were expected to easily identify both the descriptive violations and the moral misbehaviours committed by the players since the voice-over made them particularly salient. Specifically, Descriptive norm-violations happened when one of the players chose for themselves an uncommon shield colour (Figure 2A). Loyalty norm-violations happened when one of the players refused to come to the aid of the team-mate during enemies' mega-attacks, resulting in the team-mate's space-ship's destruction (Figure 2B). Fairness norm-violations happened when one of the players took for themselves all but two gems (typically the team managed to collect about 20 gems per bout prior to the division) (Figure 2C).

After each of the five internet scenarios shown (**1 refereeing familiarisation plus 4 test trials**), in a refereeing stage the child answered for each of the two players in turn: "*Did they do anything wrong?*". If a misbehaviour was identified, the child had to judge the severity of the norm-transgression ("*How bad was the player's behaviour?"*) using the 5-point smiley face scale (Figure S1 in Supplementary Information). The child was then asked to decide whether to assign a social or economic type of punishment ("*Now you can give a time-out from the game to the mean player – so that they wouldn't be allowed to play for a while – or you can take away some of their gems. Which kind of penalty do you want to give the mean player?*). Finally, the child was asked to establish the severity of the punishment (for social punishment: "*How long do you want the time out to be?*"; for economic punishment: "*How many gems do you want the mean player to lose?*", Figure 2D). Each punishment choice and consequence was accompanied by audio-visual effects, and each punishment choice was made by computer key press, to give the child the impression they were genuinely acting as referee.

At the **end of the experiment**, participants were asked whether they thought it was worse for a transgressor to receive a social or an economic type of punishment, and whether they believed they had actually refereed real internet players during the trials.

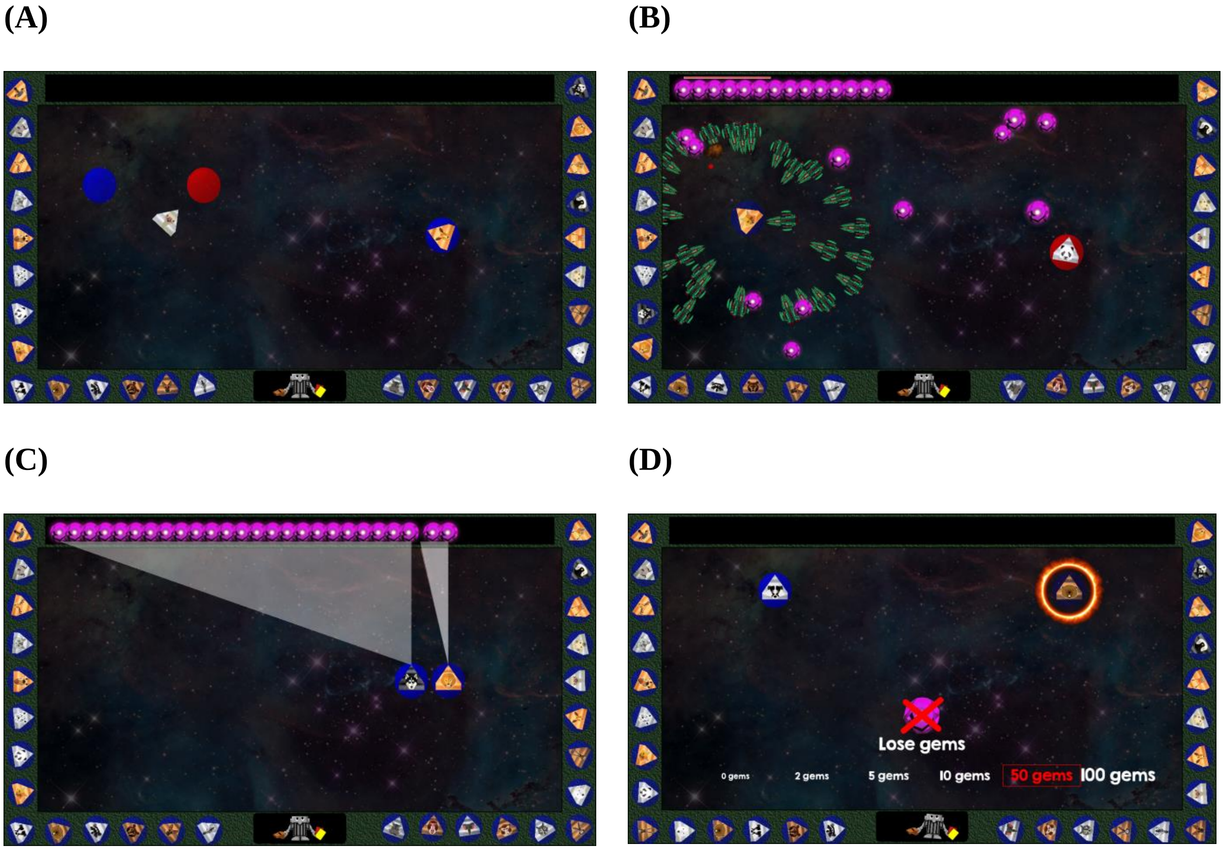

***Figure 2*. Different stages of Experiment 1 game bouts.** (A) Shield-choice stage: player Ostrich makes a descriptively non-normative choice. (B) Gem-collection stage: player Fox is under deadly threat from a Mega-attack, as disloyal player Panda ignores the situation and continues to collect gems. (C) Gem-division stage: unfair player Wolf is about to take more than their share. (D) Refereeing stage: player Beaver is about to be fined 50 gems by the participant.

Analysis Strategy and Statistics. Linear mixed-effects models were used to examine 3PP developmental patterns across moral domains and the effect of descriptive violations on 3PP severity and judgement of transgression severity, with Participants' ID included as a random factor because there were multiple data points per individual. All other IVs were included as fixed factors. Model fits were confirmed by examining diagnostic scatter plots of residuals. All analyses were conducted in the R programming environment (Version 3.6.3, R Core Team, 2020) with raw data and code available in Supplementary Information.

Results & Discussion

Preliminary analyses

Believability of the game. The majority of children (67 out of 72) expressed a belief about whether they had refereed real games. Only 37 out of these 67 children (55%) believed they had done so, implying that some children detected the deception involved. Nevertheless, there was no effect of believability on the key variables (i.e., punishment severity in Table 2; judgement of transgression severity and punishment type in Supplementary Information – section S4.4). Therefore, for the statistical analyses data is included irrespective of believability.

Punishment rate. In 279 out of the total 288 times a moral transgression was shown, children correctly recognised the violators and consequently punished them (punishment rate: 97%). Misidentification of non-violators as violators were made by 13 children, in the refereeing familiarisation (13 trials) or in the test trials (10 trials). These trials were not included in the analyses.

Main analyses

Choice of punishment types. We calculated the proportion of trials for which a punishment type was chosen in the same domain as the norm violation (i.e., economic punishment for fairness transgressions or social punishment for loyalty transgressions) to verify whether children assigned punishment types randomly or not. With only two trials in each moral domain, this proportion can only take three values (0, .5, and 1). Non-parametric analysis is therefore appropriate, so we bootstrapped (100,000 samples) confidence intervals for the proportions, along with p-values for the one-sample comparison against the null-hypothesis value of .5. For unfairness, the punishment matched the domain in 69% of trials, 95% CI [61%, 78%], $p < .001$, whereas for disloyalty the punishment matched the domain in 59% of trials which was not significant, 95% CI [50%, 69%], $p = .062$.

In order to investigate the effects of age on the tendency to make the punishment fit the crime, we also calculated an overall "Punishment Fits The Crime" (PFTC) score, as the mean of the two aforementioned proportions (i.e., proportion of unfairness trials sanctioned with economic punishment, and proportion of disloyalty trials sanctioned with social punishment) for each individual. This score did not change as a function of age, F(1,70) = 1.05, $p = .309$, $R^2 = .01$, in contrast with our prediction.

There was apparently no confound between punishment type and believed punishment severity: 20 children considered economic punishment most severe, whereas 22 considered social punishment most severe, χ^2 (1) = 0.10, $p = .758$; 25 children rated social and economic punishment as equally severe, while the remaining 2 gave no clear answer.

Children clearly made the punishment fit the crime by assigning economic costs for economic unfairness, disconfirming the *general motivation hypothesis*, according to which punishment type is entirely unrelated to transgression type (Figure 1B). However, there was no clear evidence for such a tendency for social transgressions, for which the higher level of social punishment did not reach significance. Strong support for the *specific motivation hypothesis*, according to which specific transgressions motivate specific punishments across domains (Figure 1A), is therefore also lacking. Post-hoc, we considered potential explanations for this unexpected result. For economic unfairness children might have been primed to select a form of punishment employing gems simply because gems played a salient role in the unfair scenario (*associative hypothesis*; Figure 1C). Alternatively, children's 3PP behaviour might have been additionally motivated by inequality aversion, with economic costs imposed not only to punish but also to correct unjust resource distributions (*general motivation plus equalisation hypothesis*; Figure 1D). Children of this age are indeed averse to economic inequality in third-party contexts (Shaw & Olson, 2012). The obtained results are consistent with both the associative hypothesis and the general motivation plus equalisation hypothesis because they both postulate a specific mechanism, related to gems, that causes the punishment to fit the crime for economic but not social transgressions. To distinguish these possibilities a follow-up experiment was designed (see Experiment 2).

Developmental pattern of punishment severity across moral domains. Linear mixed-effects analyses revealed that children's 3PP severity was predicted by age, moral domain of the transgression and the interaction between age and domain, while controlling for judgements of transgression severity (Table 2). Specifically, acts of unfairness were punished more severely (M = 4.44, SD = 1.09) than acts of disloyalty (M = 4.23, SD = 1.23). On average, younger children were more punitive than older children. However, this downward developmental pattern occurred only in cases of disloyalty, whereas 3PP severity remained stable across ages in cases of unfairness (Figure 3). These results were at odds with previous research analysing 3PP rates across development and are discussed after a replication attempt in Experiment 2.

***Table 2*. Modulating factors of punishment severity in Experiment 1.**

Factor	***b***	***β***	**95% CI for *β***	**χ^2**	***p***
Judgement of transgression severity	-.30	-.28	-.39, -.17	22.29	< .001 ***
Age	-.21	-.33	-.51, -.15	13.06	.001 ***
Gender	.13	.11	-.20, .43	0.49	.483
Believability	.08	.07	-.24, .38	0.20	.654
Moral domain	.19	.17	.00, .33	9.86	.007 **
Descriptivity	.03	.03	-.14, .19	0.10	.748
Age x Moral domain	.14	.21	.04, .38	6.09	.014 *

Note: * $p \leq .050$. ** $p \leq .010$. *** $p \leq .001$. For binary variables, the following categories are coded as 1 (and the others as 0): gender male, believed to be real, domain of unfairness, and descriptively uncommon choice. Raw model coefficients b are standardised to produce β and associated 95% confidence interval by normalising by standard deviation of the dependent variable in all cases and by the standard deviation of the predicting factor only when it is not categorical (age and judgement of transgression severity), meaning categorical β (gender, believability, moral domain, and descriptivity) is analogous to Cohen's *d*.

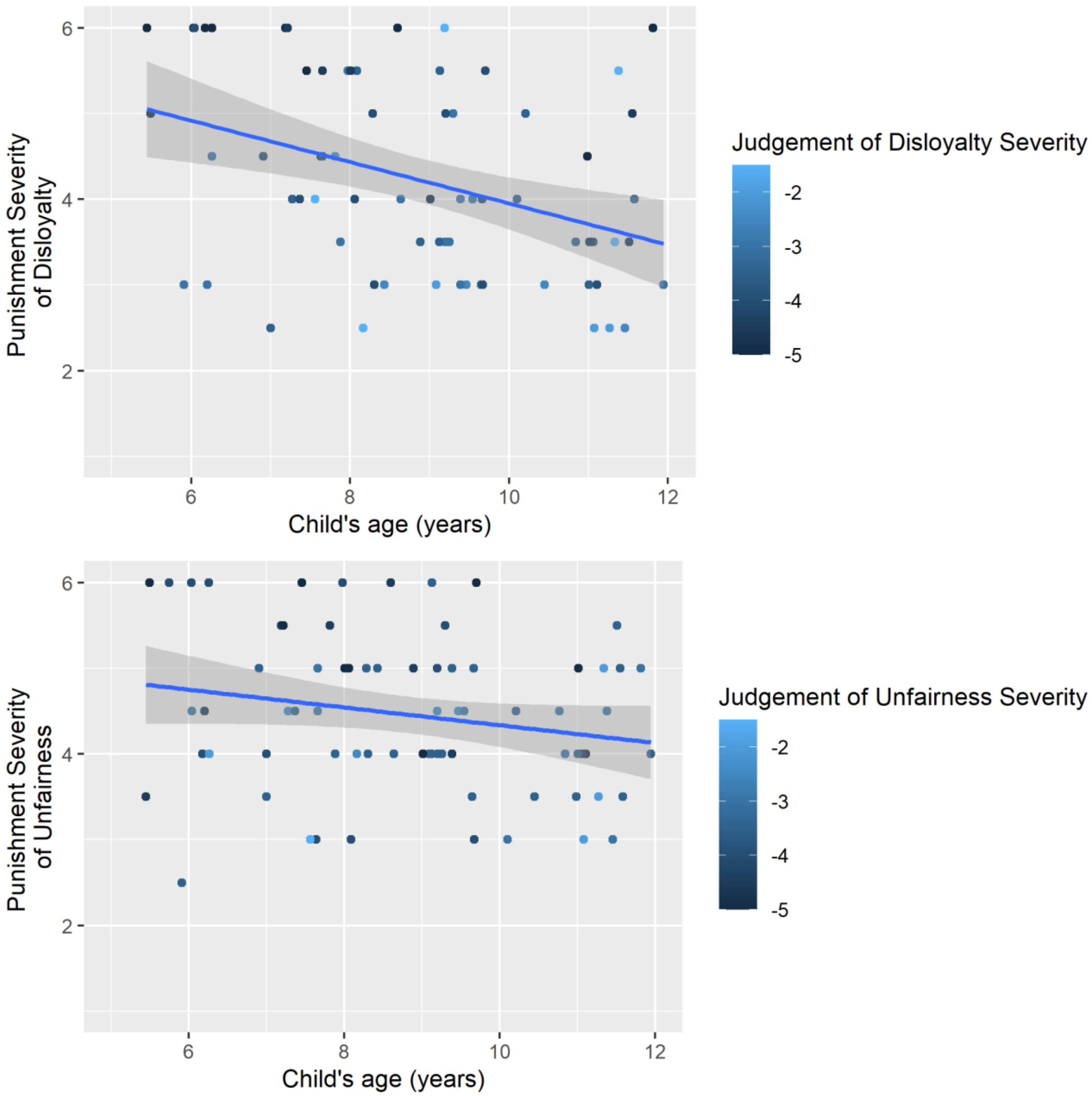

***Figure 3.* Developmental pattern of punishment severity across moral domains (disloyalty vs unfairness) in Experiment 1, with reference to judgement of transgression severity.** 95% CI of the regression line is shown.

Effects of descriptive norm violations. As shown in Table 2, descriptivity was not a predictor of 3PP severity, and the effect size confidence intervals indicate that any undetected effect is small. Further details and discussion of this result is included in Supplementary

Information (sections S4.4 and S6.1).

Experiment 2

Experiment 2 was intended to resolve the uncertainty regarding the reasons for choice of punishment types in Experiment 1; to verify whether the developmental patterns of 3PP severity were replicable; and to investigate two new issues: potential audience effects, and children's enjoyment of enactment of punishment.

Why did the punishment fit the crime for unfairness only?

Experiment 1 demonstrated economic punishment to be preferentially allocated in response to unfairness, but did not find clear evidence that social transgressions were matched with social punishment. This was most consistent with neither of the two originally proposed hypotheses, but rather with an associative explanation, or a general punishment motivation in which equalisation motives also influence behaviour (Table 3). To distinguish between these new alternative hypotheses, the transgressions were modified so that gems were made salient in the disloyal rather than in the unfair scenario, while punishment types remained unchanged (an economic punishment of a gem fine, or a social punishment of a ban). Because gems were now associated with loyalty rather than fairness transgressions, the *associative hypothesis* predicts that the economic punishment of a gem fine would now be associated with loyalty rather than fairness transgressions. In contrast, the *general motivation plus equalisation hypothesis* predicts no preference for either type of punishment in either condition, since the unfairness now concerned a different resource (bombs) that could no longer be equalised by a gem fine (Table 3).

***Table 3.* Predicted punishment preference results for each condition according to different hypotheses, plus observed results.**

Condition	**Specific**	**General**	**Associative**	**General plus equalisation**	**Observed results**
	Detection of violation within specific domain motivates punishment within domain (Fig. 1A)	Detection of violation of any domain motivates general punishment behaviour (Fig. 1B)	Salient element of transgression primes punishment involving same element (Fig. 1C)	Detection of violation of any domain motivates general punishment behaviour but equalisation motives can modify behaviour (Fig. 1D)	
Exp. 1 Disloyalty transgression	Social punishment	No punitive preference	No punitive preference	No punitive preference	No punitive preference
Exp. 1 Unfairness transgression	Economic punishment	No punitive preference	Economic punishment[b]	Economic punishment[a]	Economic punishment
Exp. 2 Disloyalty transgression	Social punishment	No punitive preference	Economic punishment[c]	No punitive preference	No punitive preference
Exp. 2 Unfairness transgression	Economic punishment	No punitive preference	No punitive preference	No punitive preference	No punitive preference

Notes:

[a] Because economic punishment (fining of gems) can help to equalise the unfair distribution of gems that motivates the punishment.

[b] Because economic punishment (fining of gems) could be primed by the featuring of gems in the transgression (unfair gem distribution).

[c] Because economic punishment (fining of gems) could be primed by the featuring of gems in the transgression (betrayal at the mega-gem).

Audience effects on moral behaviour and judgements

Audience effects – namely, behavioural changes induced by the presence of an audience or cues of observation – are known to affect punishment behaviour in adults (Kurzban et al., 2007; Piazza

& Bering, 2008). We therefore manipulated a collection of audience cues – presence or absence of a commentator and other players observing over the internet, and the attention of the experimenter – with the prediction that children would enact more severe 3PP against norm violators, and express more severe judgments about transgressions, in the Audience condition. Results of this investigation were somewhat inconclusive and further introduction and discussion of the issue is therefore provided in Supplementary Information (section S5).

Affective states involved in punishment

3PP is typically associated with negative emotions such as moral outrage and anger in response to transgressions. However, although the experience of negative emotions appears to motivate 3PP decisions in adults (Buckholtz & Marois, 2012; Gummerum, Van Dillen, Van Dijk, & López-Pérez, 2016; Lotz, Okimoto, Schlösser, & Fetchenhauer, 2011), evidence suggests this is not the case in children or adolescents (Gummerum et al., 2019). Whereas these studies have investigated the emotional antecedents to 3PP, the understanding of the emotional consequences of carrying out an act of 3PP is still incomplete. To our knowledge there are no studies of young children on this topic, and the only experimental evidence of affective correlates with 3PP in the adult literature has produced rather mixed results.

Neuroscientific studies employing dictator game and fMRI methodology have suggested that enacting 3PP is intrinsically rewarding for adult punishers. For example, after a dictator proposed an unfair offer, both second- and third-party punishers of the dictator showed stronger activation in the striatum (a brain area implicated in reward) in comparison to people who decided not to punish, although such activation was stronger in second-party punishers than in third-party punishers (Strobel et al., 2011).

Findings regarding punishers' reported satisfaction from psychological experiments are not straightforwardly reconcilable with this, however. Carlsmith, Wilson, & Gilbert (2008) carried out a public goods game where a pool of participants were informed they had all been victims of the uncooperative behaviour of a single free rider (2PP and 3PP were confounded). Punishing did have an effect on people's feelings, but in the opposite direction to expected: punishers felt worse than people who had not been given a possibility to punish. Those who simply forecasted how punishment would feel if they did punish anticipated feeling better than punishers actually did. Finally, 10 minutes after the game, punishers reported ruminating about the free rider significantly more than non-punishers.

Following Carlsmith et al.'s (2008) findings that revenge is not as "sweet" as commonly believed, experimental efforts focused on the conditions in which 2PP could be satisfying. In an experiment analysing avengers' satisfaction in relation to the reaction of the punished wrongdoer, it was found that avengers seeing a wrongdoer suffer had comparable satisfaction levels to those who decided not to punish the wrongdoer. Further, punishers who saw the wrongdoer evidence understanding and contrition in response to punishment experienced an increase in satisfaction (Funk et al., 2014; Gollwitzer, Meder, & Schmitt, 2011).

Regarding potential punishment motivations, it has been theorised that deterrence-motivated people employ punishment to teach a lesson to wrongdoers in order to deter future norm violations (forward-looking motivation), whereas retribution-motivated people use punishment because they derive, or at least expect to derive, satisfaction from inflicting damage to wrongdoers (backward-looking motivation). To provide experimental support for these conceptualisations, Crockett et al. (2014) allowed participants to pay an economic cost to sanction wrongdoers in two conditions: an open punishment condition in which wrongdoers learned that they had been punished for their transgression, argued to elicit deterrence motivations; and a hidden punishment condition in which the wrongdoer was made to believe their resource loss was due to chance rather than punishment, argued to elicit retribution motivations. Participants in the hidden punishment condition sanctioned the wrongdoer almost as frequently as in the open punishment condition. Thus, people experience satisfaction from enacting costly punishment even when there is no possibility that by punishing they could teach somebody a lesson. When asked to report their motivations to punish, people's explanations did not correspond with their behaviour as their endorsement of deterrence motivations far exceeded that of retribution motivations (Carlsmith et al., 2002).

Drawing on the experimental designs employed by Carlsmith et al. (2008), Gollwitzer et al. (2011) and Funk et al. (2014), we compared reported enjoyment levels when children were informed that they were really punishing transgressors (real punishment condition) or that they were simply sending a warning (warning condition) or that they were pretending to punish (pretend condition). Although the adult literature about punishment-related affective states is equivocal, we predicted that children would enjoy enacting punishment, as vengeance-driven retribution (Crockett et al., 2014) seems a more plausible motivation for their punishment, given that deterrence is a more cognitively demanding forward-looking motivation, and in adolescents 3PP has in fact been linked to positive affect (Hao, Yang, & Wang, 2016). Specifically, we hypothesised that children who believed they allocated actual punishment would report higher enjoyment than children who believed they were just pretending to punish. Intermediate levels of enjoyment were instead predicted for children who believed they sent warning messages to misbehaving players.

Method

Sample. Participants were 80 primary school-aged children (*mean age*: 7.91 years; *SD* = 1.62 years; *age range*: from 5.27 years to 11.56 years; 23 females and 57 males) tested in a diverse range of settings (two primary schools, three science fairs and at lab visits), but the whole testing phase took place in the same city as in Experiment 1, from December 2017 to April 2018. Power analyses were not performed because of the lack of previous data on which to base effect size expectations for the novel hypotheses, so we allowed logistical constraints to determine effect sizes.

Forty-three out of 80 caregivers (18 fathers; 20 mothers; 5 grandmothers) partially or fully completed a socio-demographic questionnaire, indicating that Experiment 2's sample came mostly from a middle-class background (the median yearly family income was £70,000; 3 out of 43 respondents preferred not to declare) with a high education level (84% of the respondents had at least a Bachelor's degree), and was predominantly British (caregivers' nationality: 38 British, 5 non-British). Data on racial identity was not systematically collected, but the sample was predominantly white.

Design. We adopted a 2x2x3 mixed design in which the factors were: *type of moral transgression* (2 within-subject levels: fairness transgression; loyalty transgression); *audience* (2 within-subject levels: present; absent); *punishment opportunity* (3 between-subject levels: real; warning; pretend), see Table 1.

We ran one trial in each of the within-subject factor combinations, for a total of four test trials. Counterbalancing was as for Experiment 1, but with audience presence or absence manipulated in place of descriptive-norm violation or conformity (see Supplementary Information – Table S3).

The dependent variables measured were: *judgement of transgression severity* (6 ordinal levels from "very bad" to "neither bad nor good", Figure S4 in Supplementary Information); *type of punishment* (3 categorical levels: gem fine, a ban, or neither of them, differently from Experiment 1, see Figure 4); *severity of punishment* (6 ordinal levels as in Experiment 1); *affective state in* *enacting punishment* (11 ordinal levels from “very bad” to “very good”, Figure S4 in Supplementary Information).

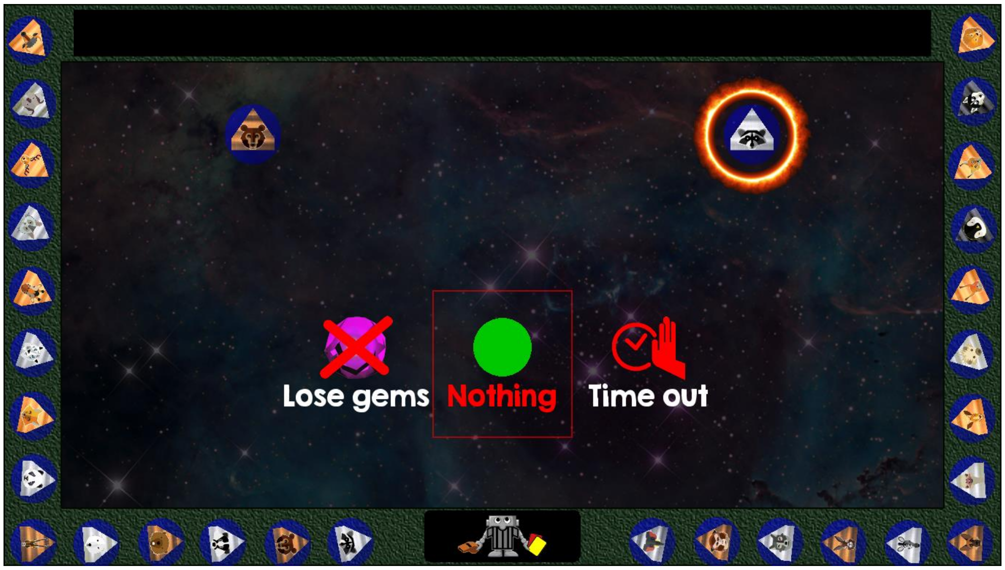

***Figure 4*. Types of punishment in Experiment 2.** Punishment severity options are the same as the ones used for Experiment 1. As a consequence, children have two possibilities to express their desire not to punish the transgressor: when they are asked to choose the type of punishment, they can select “Neither”. Should they choose either “Time out” or “Lose gems”, they can then select the no-punishment option (respectively, 0 minutes or 0 gems).

Procedure. The procedure of Experiment 2 closely resembled that of Experiment 1, thus this section describes only differences. There was no shield-choice stage and all players were automatically assigned blue shields. Game bouts still contained a gem collection stage and a resource division stage, but rather than a gem division stage after the gem collection stage, there was a bomb division stage before the gem collection stage. During the collection stage, two types of gems could appear: normal sized-gems (like in Experiment 1) and mega-gems each containing 8 normal sized-gems. The collection of the mega-gem was a cooperative task inspired by the string-pulling task (see e.g. Marshall-Pescini, Basin, & Range, 2018). For the mega-gem to be collected, both players had to attach to it. If instead only one player attached to the mega-gem, they would remain trapped, unable to protect themselves from enemies’ attacks. During **playing**

familiarisation, a loyalty norm was illustrated when the experimenter, once the child had attached to the mega-gem, cooperated with them by attaching to it too (during the third and fourth bout). There were no mega-attacks.

In the **four test trials** the live-stream commentator was now also visible as a thumbnail on the screen, to emphasise that the game was observed (Figure 5A). Loyalty violations happened when one of the players refused to cooperate with the team-mate in the mega-gem collection, thus leaving the team-mate trapped on the mega-gem, incapable of defending themselves from enemies' attacks (Figure 5A). Fairness violations happened when one of the players took for themselves more bombs than an equal share (8/10 or 9/10 in the two trials).

According to the punishment-opportunity condition children were assigned to, the purpose of the refereeing activity was framed differently in the punishment stage. Children were told they could: enact real punishment against the wrongdoers; or warn wrongdoers about possible future punishment; or just pretend to allocate punishment (see script in Supplementary Information – section S3.5 for further details about the framing).

Regarding the audience manipulation in the test trials, a range of different cues of observation were included. In the Audience condition the frame outside the game arena was full of player avatars, with animations indicating attention paid to what was happening in the arena, including the refereeing. Moreover, the stage in which the child could judge and punish the transgressors was introduced by the live-streamer with comments such as: "*Let's watch the referee making their decision*" or "*Let's see what the referee thinks*". Notably, the live-streamer remained in sight during the whole judgement/punishment phase, with the gaze directed at the refereeing child. Also, the experimenter appeared concentrated on the child's decisions. Instead, in the No Audience condition the frame around the arena was empty (i.e., no avatars formed a public) and the live-streamer, once finished commenting on the transgressions, disappeared from the screen either because of a fake internet connection problem or by pretending to move away from his computer after being called by someone, and thus could not have observed the punishment choices (Figure 5B). In order to further minimise observability cues, also the experimenter looked away from the screen, pretending to write something on a piece of paper.

At the **end of the experiment**, each child was questioned about the affective states they experienced while playing ("*How has it been playing the game with me?*") and punishing ("*So when you chose time-out or losing gems, how did it make you feel?*") by making reference to the 11-point smiley face scale, the same that participants had to use to evaluate players' transgression severity. As well as the same believability check question as previously put in Experiment 1, we also verified whether children remembered the punishment-opportunity condition they had been assigned to (real punishment; warning about future punishment; pretend punishment) by describing each and asking which applied. Finally, for exploratory purposes we asked the children whether they regretted their punishment decisions, whether they would make the same decisions and, if not, what they would do differently.

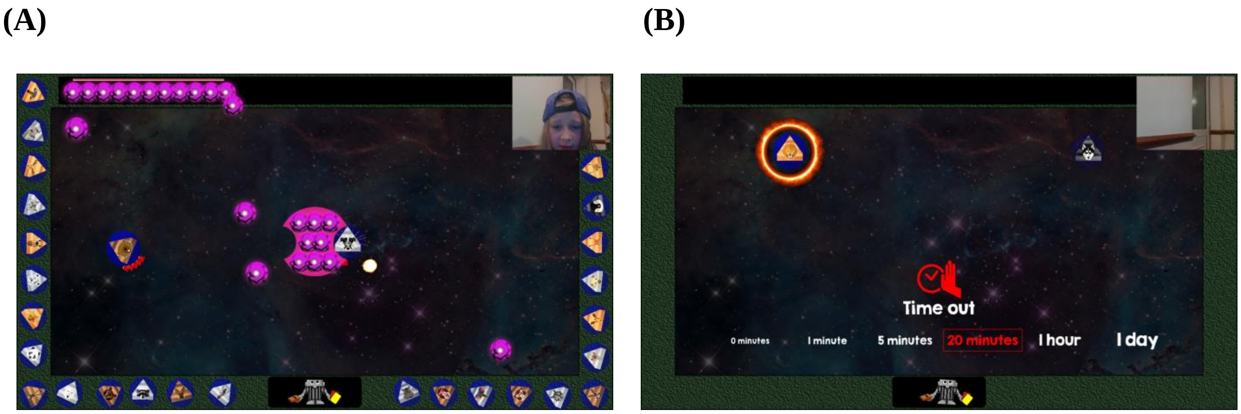

Figure 5. **Experiment 2 game bouts stages with differences to Experiment 1.** (A) Gem-collection stage: player Badger is stuck on the Mega-gem and taking damage from enemies, as disloyal player Beaver refuses to release them by also attaching to the Mega-gem to collect the gems, and the thumbnailed live-streamer observes and commentates. The authors received signed

consent for the child’s likenesses to be published in this article. (B) Referee stage: the participant is about to assign a 20-minute ban to player Lion, in the No Audience condition – there are no observing player-avatars and the live-streamer has just left.

Results & Discussion

Preliminary analyses

Believability of the game. Possibly because an apparently real live-streamer was now present on screen, commenting the players’ actions, believability apparently increased: all but one of the 80 children expressed clear beliefs, with 53 out of the 79 children (67%) believing they had refereed actual internet players during the test trials. As in Experiment 1, there was no effect of believability on the key variables (i.e., punishment severity in Table 4; judgement of transgression severity, punishment type and punishment enjoyment in Supplementary Information – section S5.4), therefore for the statistical analyses data is included regardless of believability.

Punishment opportunity manipulation check. The percentage of participants that correctly remembered the outcome of their punishment-related choices on the transgressors was 67% among children informed they were really punishing, 89% among children informed they were warning players about future punishment, and 81% among those informed they were pretending to punish.

Punishment rate. When actual transgressions were shown, in 304 out of 320 test trials (95%) children correctly identified the violators. Of these 304 trials, children chose not to punish in only 27 cases, therefore the punishment rate in Experiment 2 remained high (87%). Misidentifications of non-violators as violators were made by 2 children in the refereeing familiarisation (in one trial each) and 3 children in the test trials (in one trial each). These trials were not included in the analyses.

Main analyses

Choice of punishment types. The analysis was the same as that in Experiment 1, with proportions of trials with the punishment domain fitting the transgression domain calculated. For unfairness, the punishment domain matched the transgression domain in 51% of trials, 95% CI [42%, 60%], *p* = .941, and in disloyalty trials, the punishment domain matched the transgression domain in 42% of trials, 95% CI [33%, 50%], *p* = .057 – in other words there was no significant relations between transgression and punishment domains.

We have seen that the results of Experiment 1 were not fully in accordance with either the general or specific motivation hypotheses. The lack of a significant association between gem-related disloyalty and gem fines in Experiment 2 also runs counter to the *associative model*, according to which the preference would be for punishment that is connected to salient but superficial features of the transgression. Thus, the combined results of Experiments 1 and 2 render the *general motivation plus equalisation hypothesis* most plausible (Table 3). This suggests that children's motive to enact 3PP is not specifically related to the moral domain of the transgression; however their punishment behaviour is further modified by resource equalisation concerns. These concerns seem to lead children to select the type of punishment allowing them not only to impose a cost on the transgressor but also to equalise – when possible – the resource imbalance between the victim and transgressor. Further research will be needed, however, to confidently discard the *associative model*, as well as to investigate other potential cognitive mechanisms guiding children's choices in terms of punishment types.

Finally, in order to investigate the effects of age on the tendency to make the punishment fit the crime, we calculated again an overall "Punishment Fits The Crime" (PFTC) score, defined as in Experiment 1 as the mean of the proportion of unfairness trials punished economically and the proportion of disloyalty trials punished socially. This score did not change as a function of age F(1,75) = 0.01, p = .906, $R^2 < .001$, confirming the result of Experiment 1.

Developmental pattern of punishment severity across moral domains. Linear-mixed effects analyses revealed that children's 3PP severity was significantly predicted by age, but not by moral domain or by the interaction between age and domain, while controlling for judgements of transgression severity (Table 4). Therefore, in contrast with Experiment 1, where 3PP severity decreased with age only for cases of disloyalty, 3PP severity decreased with increasing age in cases of unfairness and disloyalty alike. Moreover, 3PP severity for acts of disloyalty (M = 4.47, SD = 1.34) was comparable to that for acts of unfairness (M = 4.31, SD = 1.44), see Figure 6.

The majority of previous literature focussed on children's 3PP rates (i.e., probability to engage vs not engage in punishment) instead of 3PP severity, and showed that 3PP rates increase rather than decrease with age (Jordan et al., 2014; McAuliffe et al., 2015; Salali et al., 2015). Therefore, the finding that, unless punishment can be used as an equalisation tool (see more detailed explanation in the General Discussion), 3PP severity is negatively predicted by age was somewhat unexpected. It is thus plausible that 3PP rates and severity are governed by different cognitive underpinnings, following different developmental patterns. However, this remains a speculative hypothesis that will need further research as the present experimental paradigm had not been designed to investigate differences between 3PP rates and severity in detail.

Although the finding that 3PP severity decreases with age had not been anticipated, it is consistent with research highlighting that children and adolescents are more severe third-party punishers than adults (Gummerum et al., 2009; Hao et al., 2016). Hao et al. suggested that decreases in 3PP severity are linked to emotional development, and in line with this we propose that the observed decrease with age of 3PP severity is possibly correlated with some components of emotion experience. Indeed, self-reported emotion ratings and activity of brain regions such as amygdala, posterior cingulate and mPFC have both been found to be associated with the severity of punishment allocated to the transgressor in adults (Buckholtz & Marois, 2012). Other explanations for this development remain plausible and further work is necessary to investigate how developing affective and cognitive processes influence children's developing 3PP behaviour.

***Table 4*. Modulating factors of punishment severity in Experiment 2.**

Factor	***b***	***β***	**95% CI for *β***	χ^2	***p***
Judgement of transgression severity	-.25	-.24	-.34, -.13	18.49	<.001 ***
Age	-.24	-.27	-.44, -.10	20.95	<.001 ***
Gender	.47	.34	-.01, .68	3.61	.057
Believability	-.38	-.27	-.60, .05	2.71	.100
Moral domain	-.01	-.01	-.17, .15	3.25	.197
Audience	.07	.05	-.11, .21	0.33	.563
Age x Moral domain	-.13	-.15	-.31, .01	3.24	.072
Punishment opportunity				1.30	.521
Actual vs. pretend punishment	-.07	-.05	-.43, .33		
Warning vs. pretend punishment	.22	.16	-.22, .54		

Note: * $p \leq .050$. ** $p \leq .010$. *** $p \leq .001$. Category coding, unstandardised (*b*) and standardised (*β*) regression coefficients with associated 95% confidence interval are the same as for Table 2, with the addition that audience presence is coded as 1 and no audience as 0.

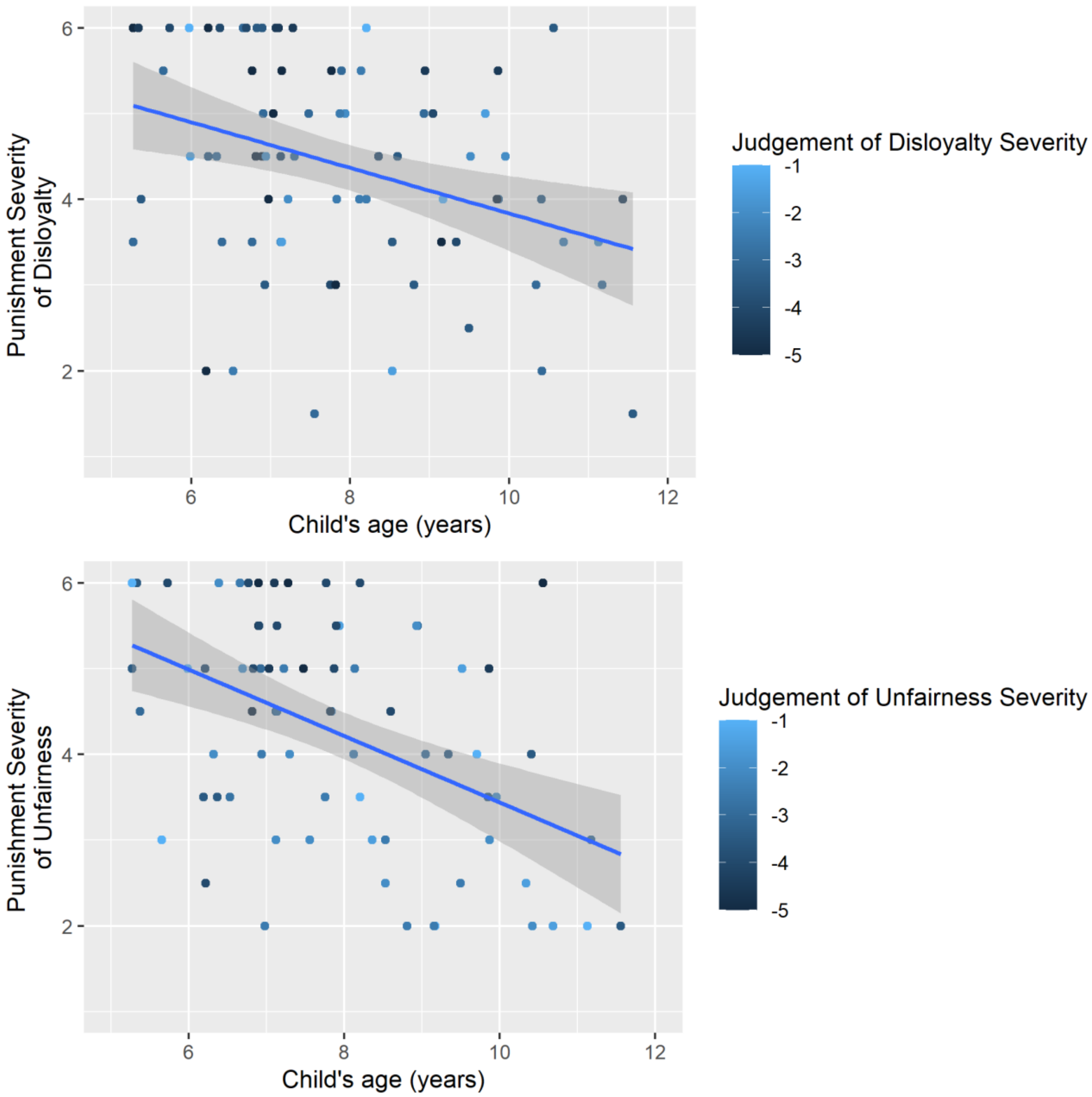

***Figure 6*. Developmental pattern of punishment severity across moral domains (disloyalty vs unfairness) in Experiment 2, with reference to judgement of transgression severity.** 95% CI of the regression line is shown.

Audience effects on moral behaviour and judgements. Children's 3PP severity was not affected by audience presence (Table 4). This null result (with confidence interval indicating any undetected effect is small) is in contrast with findings of Kurzban et al. (2007), and Piazza & Bering (2008), who observed an increase in moralistic punishment when adult participants thought their reputation was at stake. However, our audience manipulation proved to be effective in modifying children's judgements of transgression severity – see Supplementary Information (section S5.4) for further details of these results.

Affective states involved in punishment. On average children did not much enjoy making punishment-related decisions: across conditions M = 0.13, SD = 2.51, which is not significantly different from 0, t(75) = 0.46, *p* = .648, d = 0.05, 95% CI for d [-0.17, 0.27] (Figure 7). There was an association between punishment condition (real; warning; pretend) and whether the participants enjoyed punishment (enjoyment score > 0) or not (enjoyment score ≤ 0), χ^2 (2, N = 76) = 7.32, *p* = .026. Specifically, the percentage of participants that reported no enjoyment was 85% (95% CI [65%, 96%]) among children who believed they were really punishing, 58% (95% CI [37%, 77%]) among children who believed they were warning players about future punishment, and 50% (95% CI [29%, 71%]) among those who believed they were pretending to punish. Post-hoc paired comparisons (Fisher's exact tests) revealed that only the difference between real punishment and pretend punishment was significant (*p* = .044). Warning about future punishment produced a level of enjoyment intermediate between real punishment and pretend punishment, though not significantly different to either (warning-real punishment, *p* = .097; warning-pretend punishment, *p* = .777). The lack of enjoyment is unlikely to be related to idiosyncratic properties of the enjoyment scale: 95% of children reported enjoying playing the game, mean enjoyment = 4.04, SD = 1.34. Notably, the majority of children reported that they did not regret their punishment decisions (82%) and that would make the same choices again (75%). These proportions did not change depending on whether children enjoyed or did not enjoy punishment: respectively, χ^2 (2, N = 76) = .00, *p* = .987, and χ^2 (2, N = 76) = .17, *p* = .678. Among the children who declared that would not make the same choices again and explained what they would do differently, more lenient intentions (n = 8) were reported at a similar frequency than more punitive intentions (n = 6).

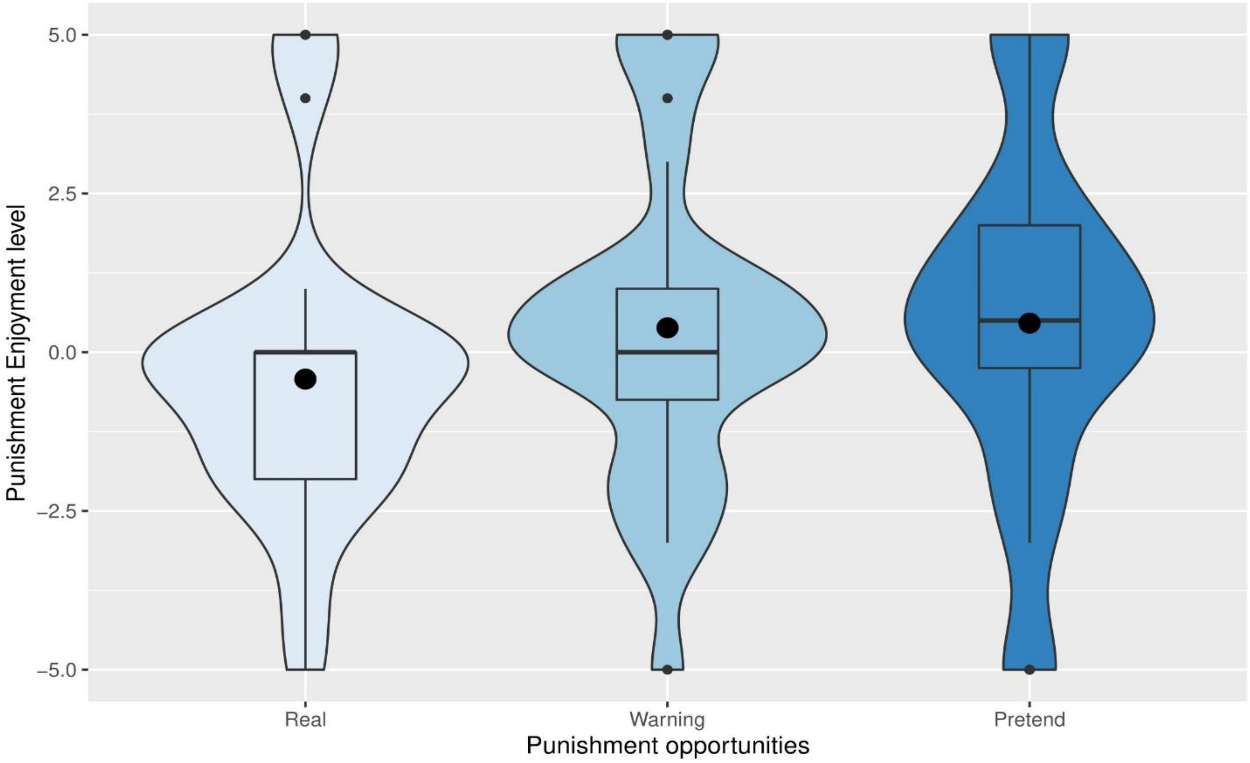

***Figure 7.* Experiment 2 punishment enjoyment by punishment opportunity condition: real; warning; pretend.** Violin plots wrapping boxplots; boxplots showing median and interquartile range, outliers, and a large dot for mean value.

Our result accords with Carlsmith et al.'s (2008) finding that punishing potentially has a negative impact on affective states, extending this result from adults (tested in a public goods game) to children (in a 3PP paradigm). Specifically, in Carlsmith et al.'s experiment punishers of free riders experienced more negative affective states than non-punishers. Furthermore, our result that lack of enjoyment was more frequent among children who believed they had allocated real over pretend punishment was particularly surprising in the light of the adolescence literature: Hao et al. (2016) found that adolescents associate 3PP with positive rather than negative affect. This lack of punishment enjoyment, accompanied by lack of regret, detected in Experiment 2 suggests that children conceptualise punishment of wrongdoers as a moral duty, something that ought to be done although it is not enjoyable. Retribution is therefore not an adequate primary explanation for the observed 3PP behaviour. In this context, it is difficult to distinguish between demand characteristics of the situation (referees are expected to punish) or deterrence motives for punishment. However, the current result suggests that especially in contexts where children punish without explicit demand characteristics (e.g., Kenward & Östh, 2015), deterrence is a more plausible motive for children’s 3PP than retribution. The extent to which children’s 3PP is motivated by implicit demand characteristics, for example a belief that adults in general approve of punishment, is an open question.

General Discussion

Our investigation has shed light on children’s 3PP by making use of an innovative and sophisticated computerised paradigm that simplified the manipulation of numerous variables embedded in a real game. In this way, we tested hypotheses of 3PP motivations, and examined the affective consequences of engaging in 3PP as well as the potential moderators of 3PP such as descriptive-to-injunctive inferences, age and audience presence.

Regarding the effect of age on 3PP, previous literature demonstrated that the odds of engaging in 3PP increased between the ages of 3 and 10 (Jordan et al., 2014; McAuliffe et al., 2015; Salali et al., 2015). With respect to 3PP severity, however, Gummerum et al. (2009) and Hao et al. (2016) found that children and adolescents were more severe punishers than adults. This is consistent with the decrease in 3PP severity between the ages of 5 and 11 we observed in both disloyalty and unfairness trials of Experiments 2, and in disloyalty (but not unfairness) trials of Experiment 1. If it is indeed generally the case that rate of 3PP increases with age but 3PP severity decreases, then it is likely that 3PP rates and severity follow distinct developmental trajectories with different cognitive underpinnings.

Moreover, our research has been the first attempt to experimentally verify whether children tend to make the punishment fit the crime in terms of moral domains. To do so we employed, across Experiments 1-2, two punishment types (social vs economic punishment) and four moral scenarios, two for each domain (unfairness: distribution of gems and distribution of bombs; disloyalty: rescue of the team-member during a mega-attack and cooperative collection of the mega-gem). The results advanced knowledge about the cognitive mechanisms used by children in punishment type decisions in two ways. Firstly, Experiments 1-2 provided evidence suggesting that there is no separation between different moral domains when it comes to the link between transgression detection and punishment motivation – there was no clear overall tendency to make the punishment fit the crime by matching social ostracism to loyalty violations and matching economic punishment to fairness violations. Secondly, we found that although the basic motive to punish therefore appears moral-domain-general, inequality aversion can substantially modify children's 3PP behaviour in terms of punishment type. Matching of the punishment to the crime was unambiguous only when the punishment could mitigate the crime (Experiment 1, gem fine for gem unfairness), which is consistent with children's well known equalisation concerns (Gummerum & Chu, 2014; Gummerum et al., 2019; Jordan et al., 2014, Smith & Warneken 2016). Further, the only condition in which punitive action could correct the results of the transgression, by equalising the unfair resource distribution, was also the only condition in which 3PP severity did not decrease with age. Although the motive to punish severely in this context is apparently generally diminishing, the lack of change in this condition is consistent with children's persistent motivations towards fairness throughout the studied age range (Shaw & Olson, 2012), if they are additionally using 3PP as an equalisation tool. This therefore additionally strengthens our *general motivation plus equalisation* account over alternative explanations.

We now turn to our most unexpected and informative result – most children showed no enjoyment of 3PP, and even warning or pretending to punish was not enjoyed by most. Nonetheless, children did not show regret for their punishment decisions and even declared they would make the same decisions again. Thus, the lack of hedonic rewards brought about by 3PP makes it unlikely for retribution to be a primary motivator of the observed 3PP, contrary to our prediction. It remains to be clarified whether lack of 3PP-related enjoyment is generalisable to other punishment contexts, or whether retribution would play a more significant role in more naturalistic settings. However, the idea that children's 3PP is not motivated by strong affective processes is consistent with findings of children's increased physiological arousal in response to transgressions prior to their engaging in 2PP but not 3PP (Gummerum et al., 2019). There are therefore two plausible explanations for the very high levels of 3PP that were observed. Children may have been motivated by deterrence, or (especially given the demand characteristics of the experiment, i.e. taking the role of a referee) children may have thought it was their moral duty to punish misbehaving players. In other words, children's punitive responses might have been at least partially motivated by the desire to conform to norms rather than to genuinely enforce moral standards of behaviour (Pedersen et al., 2018). A strong desire to conform would also be consistent with the relative lack of audience effects: perceived expectations to conform to the punishment norm might have already been close to ceiling in the No Audience condition. Importantly, note that operating according to perceived expectations is not necessarily the opposite of acting upon one's internal motivations. Over development, the one tends often to become the other – that is what norm internalisation is (but see debate about the effects of role-taking on behavioural choices in experimental settings, Levitt & List, 2007 and List, 2007).

This relates to a number of limitations that need to be acknowledged. First of all, children were likely aware they were in a testing situation rather than playing a game simply for its own sake. However, the demand characteristics in our experiments were nevertheless probably aligned with children's perceptions of adults' general expectations about 3PP, conferring some ecological validity to the situation. This claim is based on the facts that the majority of children did believe they refereed a game with real players, and that differences in behaviour were not detected in children who did not believe this. Importantly, the aim of our study was not to establish whether children punish in the absence of task demands. Our aim was rather to shed light on the cognitive and affective mechanisms governing children's 3PP behaviour. In doing so, we created some task demands to maximise the rates of 3PP and potentially the variety of 3PP responses. We thus made a trade-off decision balancing the need of a naturalistic methodology against the need of obtaining a rich repertoire of children's punitive reactions to better evaluate potential modulating factors of 3PP. As our study was designed to test our research hypotheses rather than to mimic behavioural patterns in daily life (Pisor et al., 2019), it should not be used to provide estimates of children's 3PP rates or decisions, in the real world. The frequency of 3PP behaviours, indeed, substantially differs when comparing experimental games data (like ours) to self-reports (Molho et al., 2020) or field experiments (Balafoutas et al., 2014). It is an open question the extent to which psychological mechanisms regulating 3PP are actually the same across different contexts (real life vs experiments laden with varying degrees of demand characteristics).

A second important limitation of our experimental design is that a significant minority of children did not believe the moral scenarios they were refereeing had actually occurred. However, believability rates in our experiments might be an underestimate: we asked children about the believability of the set-up in quite a conservative manner, probably bringing doubts that children had not actually experienced while they were refereeing the moral scenarios. Although reported believability did not affect the key variables we focused on, future work should aim at increasing realism of experimental settings. Believability issues, as well as the demand characteristics implicit in our study, may be tackled by employing non-supervised computerised paradigms. This would enhance the ecological validity of the methodology even further, as young children nowadays are increasingly accustomed to playing computer games by themselves. Relatedly, in order to investigate audience effects on moral judgements and 3PP we manipulated the levels of observation children were subjected to. It is worth specifying there was no condition where children certainly felt entirely unobserved, since even in the No Audience condition the experimenter was still present. Furthermore, rather than measuring 3PP propensity in terms of punishment/no-punishment binary choices, 3PP was considered on a continuum of severity. Therefore, distinct punishment severity scales were adopted, one for each punishment type. It is currently unknown whether children interpreted the time-out and fine severity scales as equivalent. However, both in Experiment 1 and 2 (where the judgement scales used were different), 3PP severity was predicted by judgements of transgression severity, adding some validity to the punishment severity scales we used. Moreover, we measured emotional consequences of 3PP engagement only explicitly. The employment of a wider set of measures (self-reported emotion ratings, skin conductance responses, facial expressions) is thus advisable to provide a more comprehensive picture of how children experience enacting 3PP.

Even though the literature on children’s punitive behaviour is growing (the number of directly relevant empirical papers has reached double digits in the last few years), there is still relatively little evidence speaking to children’s underlying motives for engaging in punishment. The finding that, at least in this context, retribution is unlikely to be an important motive for children’s 3PP was a surprising finding that highlights the importance of further investigation. Additional studies clarifying the potential roles of deterrence and conformity motivations for children's 3PP are now a priority. That multiple motivations may be involved is suggested by our conclusion that 3PP behaviour, although not generally chosen to match the specific transgression, can be modified by other related concerns such as resource equalisation. This further highlights the potential relationship between two important justice-related concerns: fairness in allocation of punishment and fairness in allocation of resources (Riedl et al. 2015; Smith & Warneken, 2016).

References

Adams, G. S., & Mullen, E. (2012). The social and psychological costs of punishing. *Behavioral and Brain Sciences*, *35*(1), 15-16. https://doi.org/10.1017/S0140525X11001142

Alter, A. L., Kernochan, J., & Darley, J. M. (2007). Transgression wrongfulness outweighs its harmfulness as a determinant of sentence severity. *Law and Human Behavior*, *31*(4), 319-335. https://doi.org/10.1007/s10979-006-9060-x

Balafoutas, L., Nikiforakis, N., & Rockenbach, B. (2014). Direct and indirect punishment among strangers in the field. *Proceedings of the National Academy of Sciences*, *111*(45), 15924-15927. https://doi.org/10.1073/pnas.1413170111

Baumard, N., André, J.-B., & Sperber, D. (2013). A mutualistic approach to morality: The evolution of fairness by partner choice. *Behavioral and Brain Sciences*, *36*(1), 59-78. http://dx.doi.org/10.1017/S0140525X11002202

Benz, M., & Meier, S. (2008). Do people behave in experiments as in the field? – evidence from donations. *Experimental Economics*, *11*(3), 268-281. http://dx.doi.org/10.1007/s10683-007-9192-y

Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms*: Cambridge University Press. https://doi.org/10.1017/CBO9780511616037

Bourrat, P., Baumard, N., & McKay, R. (2011). Surveillance cues enhance moral condemnation. *Evolutionary Psychology*, *9*(2), 147470491100900206. https://doi.org/10.1177/147470491100900206

Bregant, J., Shaw, A., & Kinzler, K. D. (2016). Intuitive jurisprudence: Early reasoning about the functions of punishment. *Journal of Empirical Legal Studies*, *13*(4), 693-717. https://doi.org/10.1111/jels.12130

Buckholtz, J. W., & Marois, R. (2012). The roots of modern justice: cognitive and neural foundations of social norms and their enforcement. *Nature Neuroscience*, *15*(5), 655. https://doi.org/10.1038/nn.3087

Carlsmith, K. M., Darley, J. M., & Robinson, P. H. (2002). Why do we punish? Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology*, *83*(2), 284. https://doi.org/10.1037/0022-3514.83.2.284

Carlsmith, K. M., Wilson, T. D., & Gilbert, D. T. (2008). The paradoxical consequences of revenge. *Journal of Personality and Social Psychology*, *95*(6), 1316. https://doi.org/10.1037/a0012165

Chernyak, N., & Sobel, D. M. (2016). “But he didn’t mean to do it”: Preschoolers correct punishments imposed on accidental transgressors. *Cognitive Development*, *39*, 13-20. https://doi.org/10.1016/j.cogdev.2016.03.002

Chudek, M., & Henrich, J. (2011). Culture–gene coevolution, norm-psychology and the emergence of human prosociality. *Trends in Cognitive Sciences*, *15*(5), 218-226. https://doi.org/10.1016/j.tics.2011.03.003

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: recycling the concept of norms to reduce littering in public places. *Journal of Personality and*

Social Psychology, 58(6), 1015. https://doi.org/10.1037/0022-3514.58.6.1015

Crockett, M. J., Özdemir, Y., & Fehr, E. (2014). The value of vengeance and the demand for deterrence. *Journal of Experimental Psychology: General, 143*(6), 2279. http://dx.doi.org/10.1037/xge0000018

Dear, K., Dutton, K., & Fox, E. (2019). Do 'watching eyes' influence antisocial behavior? A systematic review & meta-analysis. *Evolution and Human Behavior*, *40*(3), 269-280. https://doi.org/10.1016/j.evolhumbehav.2019.01.006

Delton, A. W., & Krasnow, M. M. (2017). The psychology of deterrence explains why group membership matters for third-party punishment. *Evolution and Human Behavior, 38*(6), 734-743. https://doi.org/10.1016/j.evolhumbehav.2017.07.003

Engelmann, J. M., Herrmann, E., & Tomasello, M. (2012). Five-year olds, but not chimpanzees, attempt to manage their reputations. *PLoS ONE, 7*(10), e48433. https://doi.org/10.1371/journal.pone.0048433

Ericksen, K. P., & Horton, H. (1992). "Blood Feuds": Cross-cultural variations in kin group vengeance. *Behavior Science Research, 26*(1-4), 57-85. https://doi.org/10.1177/106939719202600103

Eriksson, K., Strimling, P., & Coultas, J. C. (2015). Bidirectional associations between descriptive and injunctive norms. *Organizational Behavior and Human Decision Processes*, *129*, 59-69. https://doi.org/10.1016/j.obhdp.2014.09.011

Falk, A., & Heckman, J. J. (2009). Lab experiments are a major source of knowledge in the social sciences. S*cience*, *326*(5952), 535-538. https://doi.org/10.1126/science.1168244

Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review, 90*(4), 980-994. https://doi.org/10.1257/aer.90.4.980

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature, 415*(6868), 137-140. https://doi.org/10.1038/415137a

Fu, G., Evans, A. D., Xu, F., & Lee, K. (2012). Young children can tell strategic lies after committing a transgression. *Journal of Experimental Child Psychology, 113*(1), 147-158. https://doi.org/10.1016/j.jecp.2012.04.003

Fujii, T., Takagishi, H., Koizumi, M., & Okada, H. (2015). The effect of direct and indirect monitoring on generosity among preschoolers. *Scientific Reports, 5*, 9025. https://doi.org/10.1038/srep09025

Funk, F., McGeer, V., & Gollwitzer, M. (2014). Get the message: Punishment is satisfying if the transgressor responds to its communicative intent. *Personality and Social Psychology Bulletin, 40*(8), 986-997. https://doi.org/10.1177/0146167214533130

Gächter, S., Renner, E., & Sefton, M. (2008). The long-run benefits of punishment. *Science, 322*(5907), 1510-1510. https://doi.org/10.1126/science.1164744

Galizzi, M. M., & Navarro-Martínez, D. (2019). On the external validity of social preference games: a systematic lab-field study. *Management Science, 65*(3), 976-1002. https://doi.org/10.1287/mnsc.2017.2908

Gervais, M. M. (2017). RICH Economic games for networked relationships and communities: Development and preliminary validation in Yasawa, Fiji. *Field Methods, 29*(2), 113-129. https://doi.org/10.1177/1525822X16643709

Gollwitzer, M., Meder, M., & Schmitt, M. (2011). What gives victims satisfaction when they seek revenge? *European Journal of Social Psychology, 41*(3), 364-374. https://doi.org/10.1002/ejsp.782

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in Experimental Social Psychology* (Vol. 47, pp. 55-130): Elsevier.

Guala, F. (2012). Strong reciprocity is real, but there is no evidence that uncoordinated costly punishment sustains cooperation in the wild. *Behavioral and Brain Sciences*, *35*(1), 45. https://doi.org/10.1017/S0140525X1100166X

Gummerum, M., & Chu, M. T. (2014). Outcomes and intentions in children's, adolescents', and adults' second-and third-party punishment behavior. *Cognition*, *133*(1), 97-103. https://doi.org/10.1016/j.cognition.2014.06.001

Gummerum, M., López-Pérez, B., Van Dijk, E., & Van Dillen, L. F. (2019). When Punishment is Emotion-Driven: Children's, Adolescents', and Adults' Costly Punishment of Unfair Allocations. *Social Development*. https://doi.org/10.1111/sode.12387

Gummerum, M., Takezawa, M., & Keller, M. (2009). The influence of social category and reciprocity on adults' and children's altruistic behavior. *Evolutionary Psychology*, *7*(2), 147470490900700212. https://doi.org/10.1177/147470490900700212

Gummerum, M., Van Dillen, L. F., Van Dijk, E., & López-Pérez, B. (2016). Costly third-party interventions: The role of incidental anger and attention focus in punishment of the perpetrator and compensation of the victim. *Journal of Experimental Social Psychology, 65,* 94-104. https://doi.org/10.1016/j.jesp.2016.04.004

Haley, K. J., & Fessler, D. M. (2005). Nobody's watching?: Subtle cues affect generosity in an anonymous economic game. *Evolution and Human Behavior*, *26*(3), 245-256. https://doi.org/10.1016/j.evolhumbehav.2005.01.002

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*(50), 19931-19936.

https://doi.org/10.1073/pnas.1110306108

Hao, J., Yang, Y., & Wang, Z. (2016). Face-to-face sharing with strangers and altruistic punishment of acquaintances for strangers: Young adolescents exhibit greater altruism than adults. *Frontiers in Psychology*, *7*, 1512. https://doi.org/10.3389/fpsyg.2016.01512

Hume, D. (2000). *A treatise of human nature*. (D. F. Norton and M. J. Norton, Eds.). Oxford, UK: Clarendon Press. (Original work published 1739).

Jensen, K. (2010). Punishment and spite, the dark side of cooperation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1553), 2635-2650. https://doi.org/10.1098/rstb.2010.0146

Johnson, T., Dawes, C. T., Fowler, J. H., McElreath, R., & Smirnov, O. (2009). The role of egalitarian motives in altruistic punishment. *Economics Letters*, *102*(3), 192-194. https://doi.org/10.1016/j.econlet.2009.01.003

Jordan, J. J., & Rand, D. G. (2020). Signaling when no one is watching: A reputation heuristics account of outrage and punishment in one-shot anonymous interactions. *Journal of Personality and Social Psychology*, *118*(1), 57. https://doi.org/10.1037/pspi0000186

Jordan, J. J., Hoffman, M., Bloom, P., & Rand, D. G. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature*, *530*(7591), 473. https://doi.org/10.1038/nature16981

Jordan, J. J., McAuliffe, K., & Warneken, F. (2014). Development of in-group favoritism in children’s third-party punishment of selfishness. *Proceedings of the National Academy of Sciences*, *111*(35), 12710-12715. https://doi.org/10.1073/pnas.1402280111

Kelsey, C. M., Grossmann, T., & Vaish, A. (2018). Early reputation management: Three-year-old children are more generous following exposure to eyes. *Frontiers in Psychology*, *9*, 698. https://doi.org/10.3389/fpsyg.2018.00698

Kenward, B. (2012). Over-imitating preschoolers believe unnecessary actions are normative and enforce their performance by a third party. *Journal of Experimental Child Psychology*, *112*(2), 195-207. https://doi.org/10.1016/j.jecp.2012.02.006

Kenward, B., & Östh, T. (2012). Enactment of third-party punishment by 4-year-olds. *Frontiers in Psychology*, *3*, 373. https://doi.org/10.3389/fpsyg.2012.00373

Kenward, B., & Östh, T. (2015). Five-year-olds punish antisocial adults. *Aggressive Behavior*, *41*(5), 413-420. https://doi.org/10.1002/ab.21568

Klempka, A., & Stimson, A. (2014). Anonymous communication on the internet and trolling. *Concordia Journal of Communication Research*, *1*(1), 2.

Krasnow, M. M., Cosmides, L., Pedersen, E. J., & Tooby, J. (2012). What are punishment and reputation for?. *PLoS ONE*, *7*(9), e45662. https://doi.org/10.1371/journal.pone.0045662

Krasnow, M. M., Delton, A. W., Cosmides, L., & Tooby, J. (2016). Looking under the hood of third-party punishment reveals design for personal benefit. *Psychological Science, 27*(3), 405-418. https://doi.org/10.1177/0956797615624469

Kurzban, R., DeScioli, P., & O'Brien, E. (2007). Audience effects on moralistic punishment. *Evolution and Human Behavior*, *28*(2), 75-84. https://doi.org/10.1016/j.evolhumbehav.2006.06.001

Leimgruber, K. L., Shaw, A., Santos, L. R., & Olson, K. R. (2012). Young children are more generous when others are aware of their actions. *PLoS ONE, 7*(10), e48292. https://doi.org/10.1371/journal.pone.0048292

Levitt, S. D., & List, J. A. (2007). What do laboratory experiments measuring social preferences reveal about the real world?. *Journal of Economic Perspectives, 21*(2), 153-174. https://doi.org/10.1257/jep.21.2.153

Lieberman, D., & Linke, L. (2007). The effect of social category on third party punishment. *Evolutionary Psychology*, *5*(2), 147470490700500203. https://doi.org/10.1177/147470490700500203

List, J. A. (2007). On the interpretation of giving in dictator games. *Journal of Political Economy*, *115*(3), 482-493. http://dx.doi.org/10.1086/519249

Lotz, S., Okimoto, T. G., Schlösser, T., & Fetchenhauer, D. (2011). Punitive versus compensatory reactions to injustice: Emotional antecedents to third-party interventions. *Journal of Experimental Social Psychology, 47*(2), 477-480. https://doi.org/10.1016/j.jesp.2010.10.004

Marshall, J., Gollwitzer, A., Wynn, K., & Bloom, P. (2019). The development of corporal third-party punishment. *Cognition*, *190*, 221-229. https://doi.org/10.1016/j.cognition.2019.04.029

Marshall-Pescini, S., Basin, C., & Range, F. (2018). A task-experienced partner does not help dogs be as successful as wolves in a cooperative string-pulling task. *Scientific Reports, 8*(1), 16049. https://doi.org/10.1038/s41598-018-33771-7

McAuliffe, K., Jordan, J. J., & Warneken, F. (2015). Costly third-party punishment in young children. *Cognition*, *134*, 1-10. https://doi.org/10.1016/j.cognition.2014.08.013

McGraw, K. M. (1985). Subjective probabilities and moral judgments. *Journal of Experimental Social Psychology, 21*(6), 501-518. https://doi.org/10.1016/0022-1031(85)90022-8

Molho, C., Tybur, J. M., Van Lange, P. A., & Balliet, D. (2020). Direct and indirect punishment of norm violations in daily life. *Nature Communications*, *11*(1), 1-9. https://doi.org/10.1038/s41467-020-17286-2

Nelissen, R. M. (2008). The price you pay: Cost-dependent reputation effects of altruistic punishment. *Evolution and Human Behavior, 29*(4), 242-248. https://doi.org/10.1016/j.evolhumbehav.2008.01.001

Northover, S. B., Pedersen, W. C., Cohen, A. B., & Andrews, P. W. (2017). Artificial surveillance cues do not increase generosity: Two meta-analyses. *Evolution and Human Behavior, 38*(1), 144-153. https://doi.org/10.1016/j.evolhumbehav.2016.07.001

Okimoto, T. G., & Wenzel, M. (2011). Third-party punishment and symbolic intragroup status. *Journal of Experimental Social Psychology, 47*(4), 709-718. https://doi.org/10.1016/j.jesp.2011.02.001

Pedersen, E. J., Kurzban, R., & McCullough, M. E. (2013). Do humans really punish altruistically? A closer look. *Proceedings of the Royal Society B: Biological Sciences, 280*(1758), 20122723. https://doi.org/10.1098/rspb.2012.2723

Pedersen, E. J., McAuliffe, W. H., & McCullough, M. E. (2018). The unresponsive avenger: More evidence that disinterested third parties do not punish altruistically. *Journal of Experimental Psychology: General*, *147*(4), 514. https://doi.org/10.1037/xge0000410

Pfattheicher, S., & Keller, J. (2015). The watching eyes phenomenon: The role of a sense of being seen and public self-awareness. *European Journal of Social Psychology, 45*(5), 560-566. https://doi.org/10.1002/ejsp.2122

Piazza, J., & Bering, J. M. (2008). The effects of perceived anonymity on altruistic punishment. *Evolutionary Psychology, 6*(3), 147470490800600314. https://doi.org/10.1177/147470490800600314

Piazza, J., Bering, J. M., & Ingram, G. (2011). “Princess Alice is watching you”: Children’s belief in an invisible person inhibits cheating. *Journal of Experimental Child Psychology, 109*(3), 311-320. https://doi.org/10.1016/j.jecp.2011.02.003

Pisor, A. C., Gervais, M. M., Purzycki, B. G., & Ross, C. T. (2019). Preferences and constraints: the value of economic games for studying human behaviour. *Royal Society Open Science, 7*(6), 192090. https://doi.org/10.1098/rsos.192090

Raihani, N. J., & Bshary, R. (2012). A positive effect of flowers rather than eye images in a large-scale, cross-cultural dictator game. *Proceedings of the Royal Society B: Biological Sciences, 279*(1742), 3556-3564. https://doi.org/10.1098/rspb.2012.0758

Raihani, N. J., & Bshary, R. (2015). Third-party punishers are rewarded, but third-party helpers even more so. *Evolution*, *69*(4), 993-1003. https://doi.org/10.1111/evo.12637

Raihani, N. J., & Bshary, R. (2019). Punishment: one tool, many uses. *Evolutionary Human Sciences*, *1*. https://doi.org/10.1017/ehs.2019.12

Riedl, K., Jensen, K., Call, J., & Tomasello, M. (2012). No third-party punishment in chimpanzees. *Proceedings of the National Academy of Sciences, 109*(37), 14824–14829. https://doi.org/10.1073/pnas.1203179109

Riedl, K., Jensen, K., Call, J., & Tomasello, M. (2015). Restorative justice in Children. *Current Biology*, *25*(13), 1731-1735. https://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*, 353. https://doi.org/10.3389/fpsyg.2011.00353

Roberts, S. O., Gelman, S. A., & Ho, A. K. (2017). So it is, so it shall be: Group regularities license children's prescriptive judgments. *Cognitive Science*, *41*, 576-600. https://doi.org/10.1111/cogs.12443

Roberts, S. O., Guo, C., Ho, A. K., & Gelman, S. A. (2018). Children’s descriptive-to-prescriptive tendency replicates (and varies) cross-culturally: Evidence from China. *Journal of experimental Child Psychology*, *165*, 148-160. https://doi.org/10.1016/j.jecp.2017.03.018

Roberts, S. O., Ho, A. K., & Gelman, S. A. (2017). Group presence, category labels, and generic statements influence children to treat descriptive group regularities as prescriptive. *Journal of*

Experimental Child Psychology, *158*, 19-31. https://doi.org/10.1016/j.jecp.2016.11.013

RStudio Team. (2020). RStudio: integrated development for R. *RStudio, Inc., Boston, MA*, URL http://www.rstudio.com.

Salali, G. D., Juda, M., & Henrich, J. (2015). Transmission and development of costly punishment in children. *Evolution and Human Behavior, 36*(2), 86-94. https://doi.org/10.1016/j.evolhumbehav.2014.09.004

Schein, C., & Gray, K. (2018). The theory of dyadic morality: Reinventing moral judgment by redefining harm. *Personality and Social Psychology Review, 22*(1), 32-70. https://doi.org/10.1177/1088868317698288

Schmidt, M. F., Butler, L. P., Heinz, J., & Tomasello, M. (2016). Young children see a single action and infer a social norm: Promiscuous normativity in 3-year-olds. *Psychological Science, 27*(10), 1360-1370. https://doi.org/10.1177/0956797616661182

Shaw, A., & Olson, K. R. (2012). Children discard a resource to avoid inequity. *Journal of Experimental Psychology: General, 141*(2), 382. https://doi.org/10.1037/a0025907

Shaw, A., Montinari, N., Piovesan, M., Olson, K. R., Gino, F., & Norton, M. I. (2014). Children develop a veil of fairness. *Journal of Experimental Psychology: General*, *143*(1), 363. https://doi.org/10.1037/a0031247

Shweder, R., Much, N., Mahapatra, M., & Park, L. (1997). The “big three” of morality (autonomy, community, divinity) and the “big three” explanations of suffering. In A. Brandt & P. Rozin (Eds.), *Morality and Health* (pp. 119-169). London: Routledge.

Siegler, R. S., & Chen, Z. (2008). Differentiation and integration: Guiding principles for analyzing cognitive change. *Developmental Science, 11*(4), 433-448. https://doi.org/10.1111/j.1467-7687.2008.00689.x

Sjöblom, M., & Hamari, J. (2017). Why do people watch others play video games? An empirical study on the motivations of Twitch users. *Computers in Human Behavior*, *75*, 985-996. https://doi.org/10.1016/j.chb.2016.10.019

Smith, C. E., & Warneken, F. (2016). Children's reasoning about distributive and retributive justice across development. *Developmental Psychology*, *52*(4), 613. https://doi.org/10.1037/a0040069

Sparks, A., & Barclay, P. (2013). Eye images increase generosity, but not for long: The limited effect of a false cue. *Evolution and Human Behavior*, *34*(5), 317-322. https://doi.org/10.1016/j.evolhumbehav.2013.05.001

Strobel, A., Zimmermann, J., Schmitz, A., Reuter, M., Lis, S., Windmann, S., & Kirsch, P. (2011). Beyond revenge: neural and genetic bases of altruistic punishment. *Neuroimage*, *54*(1), 671-680. https://doi.org/10.1016/j.neuroimage.2010.07.051

Sylwester, K., Herrmann, B., & Bryson, J. J. (2013). Homo homini lupus? Explaining antisocial punishment. *Journal of Neuroscience, Psychology, and Economics*, *6*(3), 167. https://doi.org/10.1037/npe0000009

Tinbergen, N. (1963). On aims and methods of ethology. *Zeitschrift für tierpsychologie*, *20*(4), 410-433. https://doi.org/10.1111/j.1439-0310.1963.tb01161.x

Trafimow, D., Reeder, G. D., & Bilsing, L. M. (2001). Everybody is doing it: The effects of base rate information on correspondent inferences from violations of perfect and imperfect duties. *The Social Science Journal*, *38*(3), 421-433. https://doi.org/10.1016/S0362-3319(01)00132-X

Van de Vondervoort, J. W., & Hamlin, J. K. (2018). Preschoolers focus on others' intentions when forming sociomoral judgments. *Frontiers in Psychology*, *9*, 1851. https://doi.org/10.3389/fpsyg.2018.01851 van den Berg, P., Molleman, L., & Weissing, F. J. (2012). The social costs of punishment. *Behavioral and Brain Sciences*, *35*(1), 42-43. https://doi.org/10.1017/S0140525X11001348

Vogt, S., Efferson, C., Berger, J., & Fehr, E. (2015). Eye spots do not increase altruism in children. *Evolution and Human Behavior, 36*(3), 224-231. https://doi.org/10.1016/j.evolhumbehav.2014.11.007

Vrij, A., Granhag, P. A., Mann, S., & Leal, S. (2011). Lying about flying: The first experiment to detect false intent. *Psychology, Crime & Law*, *17*(7), 611-620. https://doi.org/10.1080/10683160903418213

Welch, M. R., Xu, Y., Bjarnason, T., Petee, T., O'Donnell, P., & Magro, P. (2005). "But everybody does it…": The effects of perceptions, moral pressures, and informal sanctions on tax cheating. *Sociological Spectrum*, *25*(1), 21-52. https://doi.org/10.1080/027321790500103

Wilson, D. S. & Sober, E. (1994) Reintroducing group selection to the human behavioral sciences. *Behavioral and Brain Sciences 17*:585–654. https://doi.org/10.1017/S0140525X00036104

Winking, J., & Mizer, N. (2013). Natural-field dictator game shows no altruistic giving. *Evolution and Human Behavior*, *34*(4), 288-293. https://doi.org/10.1016/j.evolhumbehav.2013.04.002

Yudkin, D. A., Van Bavel, J. J., & Rhodes, M. (2019). Young children police group members at personal cost. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0000613