CHILDREN'S THIRD-PARTY PUNISHMENT BEHAVIOUR:

THE ROLES OF DETERRENT MOTIVES, AFFECTIVE STATES AND MORAL DOMAINS

Rhea Luana Arini

Oxford Brookes University

Thesis submitted in partial fulfilment of the requirements of the award of Doctor of Philosophy

In collaboration with the Psychology Department of Los Andes University (Bogotá, Colombia)

January 2020

Table of Contents

Experiments 1-2 combined

Chapter 3: Experiment 3

Children's intent-based morality: a non-WEIRD perspective

Acknowledgments

“*It is not happiness that makes us grateful; it is gratefulness that makes us happy. […] We can be people who live gratefully: [this does not] mean that we can be grateful for everything, but we can be grateful in every given moment for […] the opportunity that is given to us.*”

(Brother David Steindl-Rast)

Three years and 4 months ago I moved from Italy to the UK, a decision that has been both tough and rewarding. It was a change in culture, language and field of research (from behavioural biology to developmental psychology) all at once, and it was pretty destabilising.

Despite this premise and my anxious attitude towards anything new, today I am a more confident and balanced person than ever before. For this to be possible, I cannot thank enough my PhD supervisors, Ben and Luci, who taught me not to fear constructive disagreements. They cultivated a working environment that was at the same time supportive and stimulating; I already know I will miss it immensely.

I want also to thank my mum, dad, sister and grandma for having always believed in my capabilities and for their moral support throughout the years and in spite of our frequent quarrels. A special acknowledgment goes to my dad for volunteering as a lab manager to help me recruit Italian participants for my last experiment!

In addition to my biological family, I am very fortunate to have a “family of choice” composed by friends, near and far. I am grateful I can confide in them, knowing I will be surely listened to, understood and encouraged.

This thesis would not exist without all the parents across the UK, Colombia and Italy who agreed for their children to participate in my experiments, and without all the children who dedicated some of their free time to my research.

I am also grateful for the opportunity I was given to start a research collaboration in Colombia with Gordon. Not only did it provide me with the working challenges I am gradually learning to welcome, but it was also an experience that transformed my sense of social identity.

I need to thank all the research assistants and volunteers at science fairs I worked with because they enriched my PhD experience and made it more enjoyable and meaningful. Thanks also to all the double coders (including my brilliant sister!) who managed to analyse my terribly noisy audio-recordings of the experiments. Thanks to all the nice people I met at conferences and seminars for the stimulating conversations.

Until a few years ago I would not have believed I could ever enjoy working out. Against all odds, I have become a regular at the gym and this has greatly contributed to my mental health and psychological resilience during the PhD. For this I can only thank my gym instructors; their help has been invaluable.

A final, ironic acknowledgment is dedicated to my immune system, which among all potential periods made me catch a flu exactly when I had to complete the revisions for this thesis.

Funding

I received financial support for the research presented in this PhD thesis through the Nigel Groome Studentship scheme (September 2016 – September 2019).

Additionally, my visiting research period in Colombia was financially supported by the Santander Research Scholarship (May – July 2018).

Declaration of contributions

I designed the four experiments included in the present PhD thesis in collaboration with my supervisors, Dr. Ben Kenward and Dr. Luci Wiggs, both from Oxford Brookes University (Oxford, UK). The conceptual development of Experiment 3 received additional input by Dr. Gordon Patrick Ingram from Los Andes University (Bogotá, Colombia).

The scripts produced in English for Experiments 1-4, as well as the translation from English to Italian for Experiment 4, were of my responsibility. The script translation from English to Spanish for Experiments 3-4 was carried out by Los Andes research assistant Juliana Bocarejo Aljure, under my supervision.

The programming of the software used for Experiments 1-3 and of the administrative control-panel interface used for Experiment 4 were performed by Dr. Ben Kenward.

Participant recruitment and data collection in the UK for Experiments 1-2 were entirely performed by me. Regarding Experiment 3, I recruited participants during my visiting period in Colombia, whereas data collection was carried out by Los Andes research assistants Juliana Bocarejo Aljure, Juan Jaccobo Garzon Martelo and John Kevin Zambrano Hidalgo. Regarding Experiment 4, the Italian sample was recruited and tested by me over the internet; the Colombian sample was recruited by Dr. Gordon Patrick Ingram and by me during my stay in Colombia, and then tested by Los Andes research assistants Juliana Bocarejo Aljure and Juan Sebastian Nassar Pereira; the British sample met face-to-face was recruited and tested by me, while the British sample met over the internet was recruited both by Brookes research assistant Marukh Mahmood and me, but tested only by Marukh Mahmood.

I conducted all the statistical analyses presented in this PhD thesis, with the guidance of Dr. Ben Kenward.

I was responsible for the whole write-up of the present thesis; critical revisions were provided by Dr. Ben Kenward and Dr. Luci Wiggs.

Publication plan

An adapted version of Chapter 2 of the present PhD thesis, including Experiments 1 and 2 combined, has been submitted for publication; the manuscript is currently under revision.

An adapted version of Chapter 4, including Experiment 4, will be submitted to a journal for publication shortly after the submission of the present PhD thesis.

An integrated version of Chapter 3, including data from Experiment 3 and additional data currently under collection, will be submitted to a journal in the near future.

A modified version of Chapters 1 and 5 combined will be submitted to a journal in the near future as a review article.

Abstract

Children engage in third-party punishment (3PP) from a young age in response to harm and fairness violations. However, several areas about children's 3PP are still un-investigated: their motivations for engaging in 3PP; the emotional consequences of enacting 3PP; and the effect of moral domains on 3PP.

In order to explore these topics, I developed two computerised paradigms: the *MegaAttack* game and the *Minecraft* Justice System. The former was used with 5- to 11-year-olds in the UK (Experiments 1-2) and Colombia (Experiment 3); the latter with British, Colombian and Italian 7- to 11-year-olds (Experiment 4). In both paradigms, as players violated different types of moral norms, children were asked to judge their behaviour and offered the opportunity to punish them. Additionally, in the *Minecraft* paradigm children could also compensate the victims.

The type of transgression children watched did not fully predict their choice of 3PP type in terms of moral domains (Experiments 1-2), but significantly affected their severity and endorsement of 3PP (Experiment 4).

Children did not appear motivated by reputational concerns, as their 3PP severity was not influenced by an audience, operationalised as cues of observation (Experiment 2) or accountability (Experiment 3).

Children's enjoyment of 3PP was generally low, although there were differences across countries (Experiments 2-3).

In Experiment 4 children enjoyed compensating more than punishing. When asked whether they endorsed deterrence or retribution as their 3PP motive, children overwhelmingly chose deterrence, irrespective of their country, age and framing manipulation received.

Reported deterrent motives, and lack of 3PP enjoyment or preference for compensation, together suggest that children, differently from adults, are not motivated by the retributive desire to see wrongdoers suffer.

Results have implications for theoretical accounts of the cognitive and affective processes involved in 3PP, methodological implications for future research avenues and, potentially, practical implications for the development of intervention studies.

(302 words)

Chapter 1: General Introduction

1.1. Normativity and mechanisms of social norm enforcement

Humans are unique among all animal species in how they regulate their social life through compliance with *social norms* (Elster, 1989). Social norms are defined as behaviours which are accepted as appropriate ways of acting towards anyone (including the self) and, therefore, generate *social expectations* (Chudek & Henrich, 2011). According to Bicchieri (2005) social expectations are to be divided into *empirical expectations* (i.e., beliefs about how people typically behave) and *normative expectations* (i.e., beliefs about how people think they ought to behave). By applying this theoretical framework it is possible to further classify social norms: *descriptive norms* are social norms motivated exclusively by empirical expectations, while *moral norms* are social norms motivated by both types of expectations. Importantly, the stronger the social expectations, the more likely people who deviate from the norms in question will incur negative sanctions such as reputation loss, overt tattling, covert gossip, shunning and punishment. On the contrary, people abiding by those norms will be rewarded with positive sanctions in the form of liking, appreciation, trust and respect (Bicchieri, 2005; Tomasello, 2014). Thus, our motivation to conform to social norms stems partly from not wanting to be disapproved of, or even punished, by others, and partly from our need to belong and to be accepted by the group (Over, 2016; Schmidt & Tomasello, 2012).

Enforcement of social norms has been thought to be of fundamental importance in order to promote cooperation (Axelrod, 1986; Fehr, Fischbacher, & Gächter, 2002), as well as to guarantee social order (Elster, 1989) and the smoothening of social interactions in general (von Rohr, Burkart, & Van Schaik, 2011). However, more recent evidence from experimental and observational studies alike (which will be discussed in this chapter in section 1.3) have brought into question the effectiveness of social norm enforcement in stabilising cooperation (for a review, see Raihani & Bshary, 2019). Thus, the purpose of social norm enforcement is perhaps more complex than was presumed.

Enforcement of social norms relies on two related mechanisms. First, individuals make judgments about others' value as social norm-followers on the basis of direct or indirect experiences (i.e., social evaluations). Second, they negatively sanction those who do not comply with social norms.

A specific case of enforcement of social norms is represented by punishment. Punishment is a sanctioning behaviour in which an individual inflicts a cost on the wrongdoer. Punishment may entail a cost for the punisher, such as emotional unease, fear of physical or social retaliation, expenditure of energetic or economical resources (Clutton-Brock & Parker, 1995; Raihani, Thornton, & Bshary, 2012). However, the extent of the cost of meting out punishment can vary greatly, depending on the context, including cases in which it is barely costly to the punisher (Pedersen, McAuliffe, & McCullough, 2018). For example, punishers do not need to fear retaliation in anonymous interactions (Piazza & Bering, 2008), or when wrongdoers are unaware that negative consequences for them are due to punishment rather than to chance (Crockett, Özdemir, & Fehr, 2014). My working definition of punishment will thus encompass also these instances in which punishers do not have to pay any physical, economic or social cost to enact punishment. Hence, in these cases the main determinant of the cost incurred by the punishers would be represented by the emotional unease they might experience in seeing the wrongdoer suffer.

Importantly, punishment can come in two main forms: second-party punishment (2PP), where the wrongdoer is punished by the victim of the norm violation, and third-party punishment (3PP), where the wrongdoer is punished by an unaffected bystander to the norm violation. Whereas second-party punishers correct the behaviour of transgressors essentially for personal benefits, third-party punishers pay a cost (of varying degree) primarily for the benefit of others (Riedl, Jensen, Call, & Tomasello, 2012). To note, the benefits victims of transgressions accrue from 2PP (by avenging themselves) or 3PP (by being avenged by someone else) are not immediate. Rather, they are delayed and conditional on a behavioural change in the punished individual towards a more cooperative disposition (Jensen, 2010; Raihani et al., 2012). Furthermore, since third-party punishers often intervene on behalf of victims they are not related to or they might not even encounter again, 3PP cannot be explained by taking into account only theories of kin selection (Hamilton, 1964) or reciprocal altruism (Trivers, 1971). For this reason 3PP, unlike 2PP, has been defined as an altruistic behaviour in biological terms, i.e. behaviour that increases the recipient's fitness to the detriment of the actor's own fitness (Hamilton, 1964). However, this does not necessarily imply that 3PP is altruistic in psychological terms, namely motivated by prosocial intentions. It is possible that 3PP is proximately motivated by the desire to see the wrongdoer suffer, and that any resulting increase in cooperation is merely an unintended, though positive, by-product (Jensen, 2010).

1.2. Developmental literature on third-party social evaluations and punishment

In order to investigate the development of normativity, it is fruitful to analyse when children start reacting to social norm violations as uninvolved third-parties on behalf of victims. Children's sense of normativity is indeed considered mature only when the enforcement of social norms is applied in an agent-neutral manner to all, rather than being parochially limited to the self or important others (Tomasello, 2014). Victims' reactions to norm transgressions will not be examined in the present PhD thesis as they are more likely (or as much likely) to arise from *personal expectations* – how victims themselves want to be treated by others – than from generalised social expectations – how they expect people to behave towards *anyone* – (De Waal, 2014; von Rohr et al., 2011).

Despite its relevance for unveiling the cognitive underpinnings of human other-regarding concerns (Jensen, Vaish, & Schmidt, 2014), the branch of developmental psychology dedicated to the study of agent-neutral normativity by means of experimental methods is still quite recent. For decades scholars have investigated the development of normativity only in verbal children, by interviewing them about their hypothetical responses to norm-violations (Hollos, Leis, & Turiel, 1986; Killen, Smetana, & Smetana, 2006; Nucci, 2001; Smetana, 1981; Smetana, Schlagman, & Adams, 1993; Stern & Peterson, 1999; Tisak & Jankowski, 1996; Tisak & Turiel, 1988). This choice has obviously excluded pre-verbal children from investigations, precluding the possibility to get a comprehensive picture of the early stages of normativity development. Moreover, interviewees recommending that a wrongdoer should "get into trouble" do not necessarily act upon their own recommendations when they themselves witness a real-life wrongdoing (FeldmanHall et al., 2012; Smith, Blake, & Harris, 2013). Therefore, in the last few years this problem has been addressed by complementing interview studies with experimental studies adopting implicit measures of children's judgements as well as observations of children's behavioural reactions to actual rather than hypothetical moral transgressions. With this regard, a summary of the most relevant experimental findings regarding children's third-party social evaluations are reported below, followed by a comprehensive literature review on the state-of-the-art in the research field of children's 3PP behaviour – which constitutes the main focus of the present PhD thesis.

Lately there has been a growing body of research on infants' reputation-related attitudes that has been based on implicit measures, such as spontaneous looking times and preferential reaching. Interestingly, infants make something like judgments from a very young age: 3- and 6-month-old infants already seem to evaluate characters differently on the basis of how they behave towards others. They demonstrate their preference for a character who has acted prosocially over one who has acted antisocially by either approaching or looking more towards the former (Hamlin, Wynn, & Bloom, 2007, 2010; Scola, Holvoet, Arciszewski, & Picard, 2015). Additionally, from 4.5 months of age infants even prefer a puppet who has acted antisocially rather than prosocially towards hindering agents (Hamlin, 2014; Hamlin, Wynn, Bloom, & Mahajan, 2011). Infants aged 12 to 18 months prefer agents who have distributed goods equally compared to agents who have performed unequal distributions (Geraci & Surian, 2011). Lastly, infants of 10 months of age expect third parties to reward fair rather than unfair distributors (Meristo & Surian, 2013), and prefer to look at the test events when they see an antisocial action performed towards an unfair rather than a fair donor (Meristo & Surian, 2014).

Once children acquire language, they start verbalising their judgements in the form of protests against those who do not follow game rules – both in explicit rule-governed games (Rakoczy, Hamann, Warneken, & Tomasello, 2010; Rakoczy, Warneken, & Tomasello, 2008, 2009; Schmidt, Rakoczy, & Tomasello, 2012) and implicit rule-governed games (Rakoczy, 2008; Wyman, Rakoczy, & Tomasello, 2009) – even when they themselves are not negatively affected by the rule breaking. From their third year of age, children have also been shown to engage in third-party interventions to defend the entitlement of others (Schmidt, Rakoczy, & Tomasello, 2013), and to prevent the destruction of someone's piece of artwork (Vaish, Missana, & Tomasello, 2011) or the theft of someone's property (Rossano, Rakoczy, & Tomasello, 2011).

Moreover, children not only make judgments about others' value as norm-followers but they also gradually take such evaluations into account in order to regulate their own interactions with these individuals, even in situations in which the norm violation is not directly experienced but just observed vicariously. For example, infants just over one year of age have been shown to reject large offers from a wrongdoer in favour of accepting smaller offers from a do-gooder (Tasimi & Wynn, 2016), and to prefer to take a toy from an individual who had allocated goods equally rather than unequally to third-parties (Lucca, Pospisil, & Sommerville, 2018). It has also been shown that children from their second year of life prefer to help a prosocial over an antisocial individual in reaching an object (Dahl, Schuck, & Campos, 2013), and choose to give favourite toys to prosocial agents and withhold toys from antisocial agents (Van de Vondervoort, Aknin, Kushnir, Slevinsky, & Hamlin, 2018). Additionally, 3- to 4.5-year-old children selectively avoid helping those who cause – or even intend to cause – others harm (Vaish, Carpenter, & Tomasello, 2010), and tend to allocate resources to prosocial rather than antisocial individuals (Hamlin et al., 2011; Kenward & Dahl, 2011). Finally, by 5 years of age children distribute more resources to those who enforce social norms compared to non-enforcers (Vaish, Herrmann, Markmann, & Tomasello, 2016).

It should be noted, however, that preference for prosocial agents may arise from increased motivation to interact with pleasant individuals, while passive avoidance of antisocial agents may arise from decreased motivation to interact with unpleasant individuals. For this reason, such reactions do not necessarily constitute a proof of children's sense of normativity. In fact, punishment of wrongdoers from a third-party perspective would be a much more conclusive evidence of children's awareness of the normative dimension of other people's actions (Kenward & Östh, 2015).

The first study in a laboratory setting to be published on children's 3PP behaviour is also the one with the youngest sample (pre-verbal children). Specifically, in the study devised by Hamlin et al. (2011) 19-month-old US toddlers were made to observe a puppet either trying in vain to open a box or dropping a ball after a short play. The puppet subsequently interacted with other two puppets, a prosocial and an antisocial character. After observing these interactions, the participants were assigned to one of two different conditions, named "Giving a Treat condition" and "Taking a Treat condition". In both conditions children were engaged in a two-alternative forced choice task. Indeed, in the Giving-a-Treat condition the experimenters distributed a treat to the children requiring them to choose whether to give it to the prosocial or the antisocial character. Instead, in the Taking-a-Treat condition the children were instructed to choose whether to take a treat away from the antisocial or the prosocial character in order to allocate it to a third character who did not have any treats. Toddlers tended to direct the Giving action towards the prosocial character and the Taking action towards the antisocial character, thus punishing the latter. In such a way, the authors demonstrated that toddlers considered more appropriate to see bad things happen to bad characters than to nice characters. Nevertheless, it remained to be clarified whether children would spontaneously punish an antisocial character when not forced to assign a negative outcome to anyone but were only provided with an opportunity to do so at their personal cost.

The same live puppet show featuring the prosocial and antisocial characters was later shown to verbal US children from 3 to 5 years of age with an important adaptation: the forced choice implicit task (i.e., taking/giving a treat to the characters) was substituted with an explicit question (i.e., "*Who should get in trouble?*"). Despite the cognitive load that especially 3-year-olds experienced, children's verbal responses consistently indicated that it was more appropriate to have punishment allocated to the antisocial rather than the prosocial character (Van de Vondervoort & Hamlin, 2017), in accordance with previous interview studies (e.g., see Killen et al., 2006).

Differently from the experiment conducted by Hamlin et al. (2011), another early study investigated young children's costly punishment behaviour without adopting a forced choice task (Robbins & Rochat, 2011). In their experiment, each child took part in a Triadic Sharing game along with two puppets. In the first rounds one puppet always generously split its coins between itself, the other puppet and the child, while the second puppet always made selfish allocations in order to keep the majority of the coins for itself. After the last round of sharing, the experimenters allowed the child to sacrifice their own coins if they wanted one of the puppets (or both) to be punished. Punishment consisted of taking away some of the puppet's coins. Children were not forced to assign a punishment to one of the two puppets: they could decide to punish one or both puppets or none. When 3- and 5-year-old US children were tested, it was shown that by 3 years of age children are willing to enact costly punishment. However, 3-year olds punished the selfish and the generous puppet indiscriminately. Only 5-year olds proved to orientate their punitive actions systematically towards the selfish puppet, thus conveying moral approval or disapproval for the puppets' actions. Interestingly, these results were not confirmed when the method was adapted for testing 5- to 6-year-old children from rural Samoa (Robbins & Rochat, 2011). Indeed, when they chose to punish, Samoan children did not express any preference for punishing the selfish over the generous puppet. Although this experiment may suggest that at least US children pay costs to prevent others from receiving unfair treatments, this design confounds 2PP and 3PP: it is possible that children punished the selfish puppet because they themselves were the victims of its misbehaviour, and not because they were motivated to intervene on behalf of the generous puppet.

Subsequently, Kenward and Östh (2012) developed an experimental protocol in which not only were children not forced to choose who to punish between two characters, but they were not even explicitly encouraged to punish, unlike in Robbins and Rochat's (2011) experiment. In Kenward and Östh's (2012) study, the target of investigation were children's punitive tendencies in response to different types of punishment demonstrations. Specifically, adult demonstrators played out stories in which a perpetrator doll made an unprovoked physically harmful attack on a victim doll (e.g., hitting, stamping on or kicking). After the attack a witness doll (third-party) intervened to enact either a consistent punishment (by targeting the perpetrator) or an inconsistent punishment (by targeting the victim). At this point, 4-year-old Swedish children were asked to retell the story, given free choice to change the witness doll's behaviour if they wanted to. It was found that, following a consistent punishment demonstration, no child modified the story. Instead, following an inconsistent punishment demonstration, the majority of children modified the story in such a way that it was now the perpetrator rather than the victim to be admonished or punished. Therefore, these results demonstrate that young children's preference for fictional scenarios in which a third-party punishes a moral norm transgressor is strong enough to overcome their propensity to imitate adults' actions, which itself has been proved to be very strong (e.g., Horner & Whiten, 2005). It still remained to be understood if children have a natural tendency to act themselves as third-party punishers against real people violating norms.

Although the studies being taken into consideration so far have the merit of indicating that punitive sentiments towards norm transgressors develop very early in human ontogeny, they have a serious limitation: all made use of fictional characters. This is problematic since it is known that children begin to understand pretence from 2.5 years of age (Walker-Andrews & Kahana-Kalman, 1999). Thus, it cannot be ruled out that older children punishing dolls or puppets realise that they are not actually punishing. For this reason in the majority of the following studies the role of the norm transgressor has been assumed by real people – children or adults depending on the specific protocol (Dixson & Kenward, *in prep;* Gummerum & Chu, 2014; Gummerum, López-Pérez, Van Dijk, & Van Dillen, 2019; Gummerum, Takezawa, & Keller, 2009; Jordan, McAuliffe, & Warneken, 2014; Kenward & Östh, 2015; Lergetporer, Angerer, Glätzle-Rützler, & Sutter, 2014; McAuliffe, Jordan, & Warneken, 2015; Salali, Juda, & Henrich, 2015).

Among the first studies employing real norm violators there are the experiments conducted by Gummerum et al. (2009), Jordan et al. (2014) and McAuliffe et al. (2015). They all adopted a three-player Dictator game paradigm whose general structure was as follows: the participants were led to believe that two people – a dictator and a recipient – had been involved in an economic game, where the dictator could decide how to split some resources (candies or coins) between themselves and the recipient, while the recipient could only accept (but not reject) the allocations. The participants were then told that their task was to determine if such allocations were acceptably fair to the recipient or not. When considered too selfish, the participants could punish the dictator by reducing their resources. However, these three experiments differed in a few important details. In McAuliffe et al. (2015) it was manipulated whether participants (5- and 6-year-old US children) had to pay a cost to punish unequal allocations and whether such inequity resulted from dictator's generosity or selfishness. Instead, in Gummerum et al. (2009) and Jordan et al. (2014) the punitive option was always costly since the participants (respectively, children of 7 and 11 years of age and adults living in Germany, and children of 6 and 8 years of age living in the US) were required to sacrifice part of their own endowment of candies or coins whenever they decided an allocation was unfair to the recipient. Moreover, in Gummerum et al.'s (2009) and Jordan et al.'s (2014) experiments the effect of group membership on children's enactment of 3PP was evaluated by manipulating whether the dictator and the recipient were in the participant's group. Specifically, in Gummerum et al. (2009) both the dictator and the recipient were either out-group or in-group members of the participant. In addition to these combinations, in Jordan et al. (2014) the dictator and the recipient could also belong to different groups, in such a way that one of them was an in-group while the other an out-group member in relation to the participant.

The study by Gummerum et al. (2009) showed that both 7- and 11-year-old children were willing to pay an economic cost to punish fairness norm violators. Their punishment severity was affected by age (i.e., children were harsher punishers than adults), but not by group membership. Conversely, adults punished unfair in-group members more severely than unfair out-group members. This makes sense if it is considered that, from punishing in-group relative to out-group members, people are more likely to reap the benefits of converting wrongdoers into cooperators (Norms-Focused Hypothesis, McAuliffe & Dunham, 2016). Another, more mechanistic explanation is that violations committed by in-group members threaten group identity and are therefore perceived as requiring higher punishment (Kerr, Hymes, Anderson, & Weathers, 1995).

In the study by Jordan et al. (2014) it was found that the rates of costly 3PP (which are distinct from 3PP severity) increased over development. Interestingly, the findings showed also that both 6- and 8-year-olds were more likely to pay to punish selfish allocations made by out-group dictators rather than in-group dictators (transgressor bias), suggesting that children, differently from adults in Gummerum et al.'s (2009) experiment, may view 3PP as a way to impose a cost upon potential competitors. Additionally, they revealed that 6-year-olds, but not 8-year-olds, were more likely to punish fairness norm violations that negatively affected in-group recipients rather than out-group recipients (victim bias). This suggests that 3PP is biased from its emergence, but that this bias partially decreased with age, meaning that children across development tend to become more impartial when enforcing norms. All the evidence gathered by Jordan et al. (2014) is thus consistent with the possibility that children's 3PP is modulated by affective preferences for in-group members (Mere Preferences Hypothesis, McAuliffe & Dunham, 2016).

Regarding the main findings of the study by McAuliffe et al. (2015), it demonstrated that, whereas 5-year-olds showed only sensitivity to the cost of punishment but not to distributional inequity, 6-year-old children were sensitive to both factors. This means that 6-year-olds punished less when doing so was personally costly, and that they punished more when the dictator's allocation was unequal – and selfish – rather than equal. Moreover, 6-year-olds have been shown to be more sensitive to inequity when it derived from selfishness rather than from generosity. In other words, they rejected unequal generous allocations more than equal allocations but less than unequal selfish allocations. Finally, McAuliffe et al.'s (2015) study is notable for having demonstrated that, when children were given the opportunity to punish a norm violation, they engaged in 3PP very frequently (in more than 80% of cases).

Another study made use of an economic paradigm, this time to assess if children tend to imitate a model's choice to enact 3PP (Salali et al., 2015). Specifically, 3- to 8-year old Canadian children were required to observe a three-player Dictator game played by two children – in the roles of dictator and recipient – and one adult model. Each participant experienced one of six different combinations of conditions and treatments: conditions were the dictator's allocations to the recipient (fair vs. unfair), while treatments were the model's decision to punish the dictator (unknown vs. no punishment vs. punishment). After having seen the dictator's proposed allocation and the model's reaction, children were given the opportunity to decide between punishing or not punishing the dictator (punishment consisted in taking away some stickers from the dictator). Punishing the dictator was costly: children had to give away one of their stickers for each punishment decision. This study showed that children, regardless of their age, are willing to pay a cost to imitate a model's action both in the case of consistent punishment (i.e., punishment of a dictator offering unfair allocations) and inconsistent punishment (i.e., punishment of dictator offering fair allocations). It is noteworthy that imitation rates increase (not decrease) with age. However, only older children imitate not-punishing for both fair and unfair allocations. These results are therefore partially in disagreement with those obtained by Kenward and Östh (2012) since these researchers had found that 4-year-old children are willing to imitate a model's action only in the case of consistent punishment, but they are able to overcome their imitative tendency when the model has enacted an inconsistent punishment. This lack of external validity could be due to the way in which the model's inconsistent punishment was operationalised in the two studies: in Kenward and Östh's (2012) study inconsistent punishment was directed towards the victim of the norm transgression (i.e., character being subjected to physical harm); in Salali et al.'s (2015) instead it was directed towards the norm-follower (i.e., dictator offering fair allocations). It is therefore likely that, when it came their turn to decide whether to imitate the model, children felt more uncomfortable in reiterating punishment on someone who already had suffered an immoral act rather than on someone who had not. Another possible explanation is that, although the model represented an authority in both cases, children believed that moral norms related to issues of harm were more authority-independent than those related to the fairness domain.

A further research group examined children's propensity to incur economic costs to enact 3PP against real norm violators. Differently from Gummerum et al. (2009), Jordan et al. (2014), McAuliffe et al. (2015) and Salali et al. (2015), in the protocol being applied by Lergetporer et al. (2014) the norm violation pertained to cooperation instead of resource distribution (fairness domain). Another important difference is that Lergetporer et al. (2014) adopted the so-called "strategy method" (Selten, 1967), whereby third-parties' punishment decisions are contingency commitments made prior to (rather than after) discovering potential violators' actual behaviour. In this experiment 7- to 11-year-old Italian children played a two-player Prisoner's dilemma game (cooperation game) in the presence of a third-party peer having the opportunity to enforce the social norm of cooperation by punishing defectors. More precisely, at the beginning of the game the third-party had to decide whether to invest a token to punish the player if this player should later defect in the game. Although players were aware they could be sanctioned by the third-party, they were not aware of the third-party's decision before choosing to defect or cooperate. Whenever the third-party decided to invest the token in punishment, the player lost all gains in case of defection. Conversely, in cases of player's cooperation, the third-party could exchange the token with a reward. By adopting such a procedure, the researchers found out that children act as third-party punishers very rarely (in less than 10% of cases), in clear contrast with what was found by McAuliffe et al. (2015), whose experiment showed high rates of 3PP (in more than 80% of cases). But this low incidence of punishment might be due to the fact that third-parties had to decide to invest their resources in punishment before – rather than after – having seen the violation of the cooperative norm. These results were then compared with a control condition in which the players played the Prisoner's dilemma game without being observed by a third-party. The comparison demonstrated that the sole threat of 3PP more than doubles players' cooperation rates (58% in the presence of third-party vs 25% in the absence of third-party). This might indicate that 3PP has evolved as a mechanism to sustain large-scale networks among unrelated strangers by discouraging violations of cooperation norms (see related evolutionary debate in section 1.3).

Other examples of the strategy method being applied to the study of children's 3PP, this time in comparison with 2PP, can be found in Gummerum & Chu (2014) and Gummerum, López-Pérez, Van Dijk, & Van Dillen (2019). Differently from Lergetporer et al. (2014), in these two experiments punishment was designed to be elicited by unfair distributions of resources rather than by lack of cooperation. Specifically, in the 3PP condition children played a three-player Dictator game in which the child was the potential third-party punisher. Each 3PP commitment required children to invest some of their economic resources to reduce the dictator's payoff. In the 2PP condition children took instead the role of the responder in a two-player Ultimatum game. The Ultimatum game is similar to a two-player Dictator game (i.e., game involving only a dictator and a recipient), but differs from it in that the responder has the possibility to either accept or reject the allocation decided by the proposer. In case of acceptance, the resources are divided according to the proposal; in case of rejection (which is a form of 2PP), both players receive nothing.

Gummerum & Chu's experiment (2014) was aimed at investigating the development of the capability to integrate outcome and intention information into punishment behaviour in a UK sample. Participants (children, adolescents and adults) were asked how they would react to a range of possible allocations made by the dictator in the 3PP condition or by the proposer in the 2PP condition. As in Lergetporer et al. (2014), participants had to commit to a punishment decision before they discovered the dictator/proposer's actual allocation. Crucially, participants knew that the dictator/proposer could decide the allocation by choosing between two fixed options: the default one was unequal to the advantage of the dictator/proposer, while the alternative one could be more unequal, equal, or advantageous to the recipient/responder. The contrast between the two options thus allowed inference regarding the dictator/proposer's intention: e.g., choosing the default unequal option when the alternative was even more unequal would be indicative of positive intentions, whereas choosing the default unequal option over the equal alternative would indicate negative intentions. It was found that 8-year-old children were sensitive only to information about the outcome of the allocations (i.e., whether they were equal or unequal), whereas adolescents took into consideration both outcomes and intentions in 2PP but not in 3PP. The integration of outcome and intention information in both 2PP and 3PP began to be detectable only from early adulthood. Measures of reaction times revealed that for adults and partly for adolescents making intent-based punishment decisions was more cognitively demanding when outcome information (i.e., dictator/proposer choosing an unequal allocation) was in contrast with intention information (i.e., dictator/proposer discarding an even more unequal alternative).

Gummerum et al. (2019) investigated the role of emotions in motivating costly 2PP and 3PP against unfairness violations, respectively in the two-player Ultimatum game and three-player Dictator game, across different age groups (children, adolescents and adults living in the UK). Different components of emotional appraisals were taken into account: one more automatic, arousal, and the other more deliberate, valence. Arousal was measured via skin conductance while valence via explicit emotion ratings. It was found that, whereas unfairness of proposer/dictator's allocations influenced both 2PP and 3PP across all age groups, the effect of emotions varied depending on several factors: self-relevance of the transgression (2PP vs 3PP), participants' age, and type of emotional appraisal considered. Specifically, 2PP was associated with both more negative emotional valence and higher emotional arousal (especially in children compared to adolescents and adults). However, valence mediated the link between unfairness and 2PP in all age groups, while arousal did so only in adults. In order to explain these age differences in the mediation analyses, Gummerum et al. (2019) referred to Dys and Malti's (2016) argument that during development emotional reactions may gradually become automatic following repeated associations between the occurrence of a moral transgression and its deliberate emotional appraisal. The "automatisation" of emotional appraisal processes would thus be indicative of moral norm internalisation. From this perspective, Gummerum et al.'s (2019) 2PP results would be explained by children and adolescents being solely reliant on deliberate emotional appraisal to prompt them to engage in 2PP. On the contrary, adults would have already automatised their emotional reactions so that arousal and valence are in line to jointly elicit 2PP. Differently from 2PP, 3PP was associated with more negative emotional valence but not higher emotional arousal. Moreover, valence mediated the link between unfairness and 3PP only in adults, while arousal in neither age group. These findings are therefore consistent with the view that automatic emotional appraisal can be elicited only in second-party contexts, when the transgression is self-relevant to the punisher (Civai, Corradi-Dell'Acqua, Gamer, & Rumiati, 2010). In third-party contexts, instead, participants may engage in deliberate emotional appraisal to the view of a victim receiving an unfair allocation. However, for their negative emotions to motivate costly 3PP further deliberate processes seem to be needed, such as affective perspective-taking with the victim. In children and adolescents the motivating role of emotions in 3PP enactment is thus prevented by their under-developed perspective-taking skills (Will, Crone, van den Bos, & Güroğlu, 2013). Conversely, when adults are confronted with transgressions as third-parties, their capability to take the victim's perspective may contribute to bridge the "self-relevance gap".

As pointed out by Kenward and Östh (2015), a limitation of the experimental studies described so far is that costs being paid by the children who decide to punish are defined in economic terms, whereas in real life such costs are frequently social and consist of a risk of retaliation from the punished individual (Janssen & Bushman, 2008). Kenward and Östh (2015) therefore developed an experiment in which 5-year-old Swedish children were shown a video representing two adults (presented as real): an antisocial individual who destroyed a gift they had just received, and a neutral individual who kept the gift. Children were then given the possibility to assign positive (normal good-tasting sweets) or negative outcomes (disgusting-tasting fake sweets) to the two adults. The experimenters manipulated whether children were led to believe they would allocate sweets to the recipient anonymously or in person. It was found that, both in the anonymous and in-person condition, neutral individuals were assigned only positive outcomes in the great majority of trials. Conversely, antisocial individuals were usually assigned negative outcomes, as long as children were told they would remain anonymous. Most children who were instead told they would make the allocation in person did not assign negative outcomes, although a minority did so. This indicates that children considered the situation as real and that, although they would have wanted to punish antisocial adults, they preferred not to run the social risk to do so in person. It is noteworthy that boys punished more frequently than girls.

When the aforementioned protocol was adapted for testing Melanesian children, it was interesting to analyse cultural comparisons. In this new study carried out by Dixson and Kenward (*in prep*), 4- to 10-year old Melanesian children had the possibility to distribute good or disgusting sweets to an antisocial or neutral adult in one of four different conditions, obtained by combining the presence or absence of economic and social costs inherent in the choice to punish. Thus, the conditions were: “free anonymous” (neither economic nor social costs), “costly anonymous” (economic cost but no social cost), “free in-person” (social cost but no economic cost) and “costly in-person” (both economic and social costs). It turned out that also Melanesian children punished more the antisocial than the neutral individual. Moreover, the punishment rate towards the antisocial individual was shown to increase with age and to decrease when it was economically costly and when meted out in person. Nevertheless, almost a third of Melanesian children punished the antisocial individual in the costly anonymous condition.

Lately, children’s punitive justice has been directly compared with restorative and distributive justice, although not in situations where children interact with actual people. The punitive-restorative justice comparison has been dealt with in a study conducted by Riedl, Jensen, Call and Tomasello (2015) where 3-year-old German children, after having witnessed resources being taken away from a doll victim by a doll transgressor, had to choose whether to give back the resources to the victim (i.e., restoration) or to make them inaccessible to the transgressor (i.e., 3PP). Neither choice was costly (at least in physical, economic or social terms) to the participants. Children proved to have a preference for restoration over 3PP, demonstrating to prioritise concern for the victim over consideration about the outcome for the transgressor. Smith and Warneken (2016) tackled instead the punitive-distributive justice comparison by examining US children's allocations of rewards (e.g., feeding pets or testing computer games) and punishments (e.g., cleaning tables) to fictional students depending on the information about their past behaviour (i.e., whether they had behaved well or not in the classroom). Both types of allocations were at no personal cost for the participants. Children of 4-5 years of age were more likely to split both rewards and punishments equally among the fictional recipients. Vice versa, older children and adults were more sensitive to issues of merit, thus assigning more rewards to students who were well-behaved and more punishments to students who misbehaved more. Interestingly, the developmental trajectory of punishment allocation decision-making was perfectly mirrored by that of reward allocation.

One of the most recently published studies about children's 3PP investigated whether punishment rates are affected by the type of punishment children can enact against norm violators. Specifically, in the study conducted by Marshall, Gollwitzer, Wynn, and Bloom (2019) 4- to 7-year old US children watched a puppet show on a laptop depicting a prosocial and an antisocial character interacting with the protagonist puppet. Through a button board children were instructed they could administer either rewards (i.e., tickles) or corporal punishment (i.e., hits with a hammer) to each character. Whereas the prosocial character was tickled more than hit, the antisocial character was not hit significantly more than tickled, thus showing no evidence of selective 3PP. This indicates that children's willingness to enact 3PP varies depending on the type of punishment at their disposal. Inflicting harm on a transgressor might thus be considered more socially inappropriate by children than withholding or taking away resources (Gummerum & Chu, 2014; Gummerum, López-Pérez, et al., 2019; Gummerum et al., 2009; Hamlin et al., 2011; Jordan et al., 2014; McAuliffe et al., 2015; Riedl et al., 2015; Robbins & Rochat, 2011; Salali et al., 2015). It is unclear, however, how to reconcile Marshall et al.'s (2019) null result with Kenward and Östh's (2015) finding that children engage in harm-inflicting 3PP at least when their anonymity is guaranteed. However, the punishment types adopted in the two studies had both a negative impact on the transgressor's physical sphere but at different levels. It is thus possible that children tested by Kenward and Östh (2015) showed a propensity to punish while those tested by Marshall et al. (2019) did not because punishment severity is arguably higher in the case of hitting someone with a hammer than in the case of giving someone bad tasting sweets. Children might be adverse to a corrective use of high severity-physical harm.

The last study published on children's 3PP was conducted by Yudkin, Van Bavel and Rhodes (2019). They tested 3- to 6-year old US children using a naturalistic method: participants were shown videos depicting two actual children – an actor and a recipient – who were drawing pictures. The recipient had to leave for a while so they asked the actor to hold their picture until they came back. Then, the actor was shown either ruining (harmful condition) or carefully holding the recipient's picture (harmless condition). Participants could enact two types of 3PP: the costly version consisted of closing a slide the actor wanted to play with – a decision that would have also prevented the punisher from going down it. The relatively non-costly version consisted in withholding a sticker from the actor. In accordance with McAuliffe et al. (2015), it was found that children punished the harmful actor more frequently when 3PP was non-costly than when it was costly. Moreover, rates of both costly and non-costly 3PP increased with age, comparably to what was observed by Dixson and Kenward (*in prep*) and Jordan et al. (2014). With regard to costly 3PP, Yudkin et al. (2019)

also manipulated participants' sense of authority (i.e., whether the punisher wore a symbol of leadership or not) and group membership (i.e., whether the children in the video belonged to a different or the same group as the punisher). Older children were not influenced by either manipulation; conversely younger children in a position of authority punished the harmful actor more frequently when they were in-group rather than out-group members. This latter result is thus in the opposite direction to what was found by Jordan et al. (2014) in children, but clearly mirrors Gummerum et al.'s (2009) finding in adults. Finally, when Yudkin et al. (2019) analysed the explanations children provided for their punitive decisions, it emerged that children's justifications that focused on the desire to see the transgressor change their behaviour and learn a lesson correlated with punishment rates. Interestingly, this raises the possibility that children are motivated by deterrence (Carlsmith, Darley, & Robinson, 2002) and pedagogical considerations (Funk, McGeer, & Gollwitzer, 2014) when they administer 3PP.

To sum up, children engage in 3PP from a very young age (Hamlin et al., 2011), even when they have to pay an economic cost (Gummerum & Chu, 2014; Gummerum, López-Pérez, et al., 2019; Gummerum et al., 2009; Jordan et al., 2014; Lergetporer et al., 2014; McAuliffe et al., 2015; Robbins & Rochat, 2011; Salali et al., 2015) or a social cost (Dixson & Kenward, *in prep;* Kenward & Östh, 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, et al., 2019; Gummerum et al., 2009; Jordan et al., 2014; McAuliffe et al., 2015; Robbins & Rochat, 2011; Salali et al., 2015) or harm (Hamlin et al., 2011; Kenward & Östh, 2012, 2015; Van de Vondervoort & Hamlin, 2017). Types of punishment investigated have mainly consisted of children withholding or taking away resources from transgressors (Gummerum & Chu, 2014; Gummerum et al., 2019; Gummerum et al., 2009; Hamlin et al., 2011; Jordan et al., 2014; McAuliffe et al., 2015; Riedl et al., 2015; Robbins & Rochat, 2011; Salali et al., 2015), or inflicting them harm (Dixson & Kenward, *in prep;* Kenward & Östh, 2015; Marshall et al., 2019). It has been demonstrated that 3PP rates in children increase in response to modelling (Salali et al., 2015) and with age (Dixson & Kenward, *in prep;* Jordan et al., 2014; McAuliffe et al., 2015; Salali et al., 2015), but that 3PP severity decreases with age (Gummerum et al., 2009). There is also indication that gender (Dixson & Kenward, *in prep;* 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 et al., 2019). Additionally, children being asked to allocate 3PP show a shift from equality-based to merit-based distributions with increasing age (Smith & Warneken, 2016), but not from outcome-based to intent-based distributions until adulthood (Gummerum & Chu, 2014). 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 are deterrence- and pedagogy-focused (Yudkin et al., 2019). Finally, the experience of negative emotions does not appear to motivate 3PP decisions in children (Gummerum, López-Pérez, et al., 2019).

From the review of children's 3PP literature just provided, it appeared that little attention has been paid so far to the development of 3PP motivations. In order to discuss this topic, I will now take an adaptationist stance, according to which cognitive mechanisms have functions shaped by selective pressures. I indeed believe that research investigations at the mechanistic level can be greatly aided by functional considerations. Shifting focus from the developmental to the evolutionary literature about 3PP will thus inform the design of the experiments for the present PhD thesis, by guiding me in generating novel proximate research questions (MacDougall-Shackleton, 2011). I anticipate here that the study of the adaptive functions of 3PP behaviour in adults (see section 1.3) has allowed me to identify the following as potential candidates of children's 3PP motivations to further explore at the proximate level: retribution (i.e., punishment as an end in itself); deterrence (i.e., punishment as a means to an end, namely deterring future transgressions); and reputation (i.e., punishment as a signalling tool).

1.3. Evolutionary explanations of third-party punishment in humans

Compared to other animal species, humans cooperate at an unprecedentedly high level, especially with non-relatives and strangers. Punishment of wrongdoers by third-parties has been traditionally indicated as a key factor in sustaining the progressive establishment of large-scale cooperative networks by discouraging violations of cooperation norms (Boyd & Richerson, 1992). 3PP seems to be a human universal as all populations tested so far – including urban, rural and nomadic populations – have shown some willingness to punish norm transgressors, albeit with significant cross-cultural variation in terms of frequency and severity (Henrich et al., 2006; Marlowe et al., 2007). However, the emergence of 3PP in our species constitutes an evolutionary conundrum: apparently, this behaviour is beneficial for the group's welfare but is detrimental for the punishers themselves, who suffer an immediate fitness disadvantage relative to their group members. As punishers do not obviously compensate their fitness costs with kin or reciprocity benefits, scientific efforts have been focused on elucidating the conditions that made it possible for 3PP to be under positive selection (Raihani & Bshary, 2019).

Computational models to account for the evolution of 3PP have hypothesised that members of a group can choose one of three alternative behavioural strategies when confronted with cooperation dilemmas. People can be *pure cooperators*, who abide by cooperation norms but never punish those who do not follow suit; *non-cooperators*, who neither adhere to cooperation norms nor punish; or *strong reciprocators*, who both comply with cooperation norms and punish norm transgressors. Agent-based simulations have demonstrated that, for stabilising cooperation in large groups of genetically unrelated individuals, strong reciprocators must be in equilibrium with cooperators and non-cooperators (Bowles & Gintis, 2004; Gintis, 2000).

According to the *cultural group selection hypothesis*, proximate preferences for strong reciprocity can spread in the population via social learning mechanisms, such as conformist transmission (Henrich & Boyd, 2001), particularly under conditions of strong between-group competition. Thus, groups with a more pronounced propensity to punish violations of cooperation norms would have a competitive advantage by virtue of their resulting higher levels of within-group cooperation. It is indeed likely that in our ancestral past more cooperative groups were at lower risk of extinction because they were more capable of managing the storage of common resources, to jointly patrol their territory, to organise hunting and warfare (Boyd, Gintis, Bowles, & Richerson, 2003). In such an evolutionary scenario, third-party punishers would derive inclusive fitness benefits (i.e., the success of an individual in passing on its genes to the next generation, considering also the shared genes passed on by its kin, Hamilton, 1964) from the increased prosperity of their own group (Raihani & Bshary, 2019).

Importantly, the main assumptions of group selection accounts of 3PP are that: 1) Punishment evolved because of its capacity to induce wrongdoers to assume a more cooperative attitude in the future. 2) Punishment-mediated increase in cooperation improves the welfare of the group (Raihani & Bshary, 2019).

Classic studies of punishment employing economic games have shown that average intra-group cooperation levels, in the form of contributions to the "public good", are higher when participants are given the possibility to punish non-cooperators than when such possibility is not provided (Fehr et al., 2002; Fehr & Gachter, 2000; Fehr & Gächter, 2002). However, this does not prove that non-cooperators turn into cooperators after being subjected to punishment. In other words, these studies do not control whether increased intra-group cooperation is the result of conditional cooperation or actual punishment. A more recent study seems to point at the former rather than the latter explanation. Although counterintuitive, it was found that the possibility of being punished increased both the number of non-cooperators and the average contributions to the public good (Kirchkamp & Mill, 2018). Thus, it might be that the act of punishment itself is not effective at modifying non-cooperators' behaviour. However, the threat of punishment might reassure people who are already willing to cooperate that those who have exploitive intentions will pay a cost if they decide to actualise them (Raihani & Bshary, 2019).

Although it remains to be verified whether 3PP is able to improve intra-group cooperation rates (i.e., proportions of cooperators in a group) under certain conditions, the aforementioned studies nevertheless demonstrated the positive impact of 3PP on group's cooperation levels (i.e., average contributions of cooperators) (Fehr et al., 2002; Fehr & Gachter, 2000; Fehr & Gächter, 2002; Kirchkamp & Mill, 2018). However, growing evidence suggests that the effectiveness of 3PP even in promoting cooperation levels is context-specific. For example, the benefits of 3PP are higher when: 3PP is mild vs severe (Jiang, Perc, & Szolnoki, 2013); 3PP is enacted in wealthy high-trust societies vs low-trust ones (Balliet & Van Lange, 2013); and punished individuals are reminded of third-party punishers' contributions to the "public good" rather than of their earnings (Nikiforakis, 2010).

Moreover, critics of group selection models of 3PP have pointed out that traditional "public good" experimental paradigms ignored a key factor while evaluating group's welfare:

the risk of third-party punishers being retaliated against. Retaliation is indeed impossible in one-shot interactions, when punishers' identity is kept anonymous and group composition is modified at every round of the game (Fehr et al., 2002; Fehr & Gachter, 2000; Fehr & Gächter, 2002). Although these expedients had been adopted to rule out the possibility of reciprocity and reputation formation among participants, they compromised the ecological validity of the experiments. When participants have instead the opportunity to retaliate, 3PP does not always promote cooperation, but it can trigger an escalation of counter-punishment, detrimental for the group's welfare. Specifically, when 3PP can be avenged, people's willingness to act as third-party punishers against non-cooperators drastically drops (Balafoutas, Grechenig, & Nikiforakis, 2014). In these conditions lower 3PP rates can lead to a progressive reduction in group's cooperation levels compared to situations in which 3PP cannot be avenged (Nikiforakis, 2008; Nikiforakis & Engelmann, 2011). Under threat of retaliation group cooperation levels are thus comparable to those observed when individuals do not have any 3PP opportunities in the first place (Engelmann & Nikiforakis, 2015). Although not all experimental studies have found that mushrooming of counter-punishment is accompanied by lowering of average group's cooperation levels (Fehl, Sommerfeld, Semmann, Krambeck, & Milinski, 2012), several computational models do not support the evolution of human cooperation when non-cooperators have the chance to retaliate against third-party punishers (Hauser, Nowak, & Rand, 2014; Janssen & Bushman, 2008; Rand & Nowak, 2011).

It is also noteworthy that people assign 3PP to non-cooperators even in scenarios where there is no chance for the group to benefit from a potential change in the targets' behaviour. Not only do people enact 3PP in one-shot interactions (Fehr et al., 2002; Fehr & Gachter, 2000; Fehr & Gächter, 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). On a proximate level of analysis, this suggests that people are willing to engage in costly 3PP purely for the sake of giving the wrongdoers their "just deserts", without any further instrumental reason (either deterrent or pedagogical). Such a behaviour would thus be in line with a ***retributive motivation*** to punish. On an ultimate level of analysis, it has been argued that the adaptive function of 3PP is not to reform non-cooperators' behaviour, but to decrease their fitness in the current generation and thus their number in the following generations. For this to be possible, there is no necessity for non-cooperators to learn any moral lesson (Crockett et al., 2014).

Interestingly, people have also been shown to engage in antisocial punishment – namely punishment of cooperators rather than non-cooperators – especially after having been themselves targets of 3PP (Herrmann, Thöni, & Gächter, 2008). This form of "do-gooder derogation" might represent a way for antisocial punishers to level out group members' payoffs so that none stands out against the others, or to increase their own payoff to establish a more dominant status within the group hierarchy (Sylwester, Herrmann, & Bryson, 2013). Relative payoff concerns can also offer an explanation for third-party punishers' sensitiveness to inequality. Indeed, those 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 proximately motivated more by envy of the wrongdoer's higher payoff than by moralistic anger towards the unfair behaviour victims had to experience (Pedersen, Kurzban, & McCullough, 2013).

All this evidence thus weakens the claim that 3PP is a tool at the service of within-group cooperation. At the most, according to this perspective, the cooperation-enhancing effects of 3PP that occur in some circumstances might just constitute an evolutionary by-product. Therefore, group selection explanations of the evolution of 3PP have been recently challenged by individual selection accounts, such as the signalling and deterrence hypotheses (Krasnow, Delton, Cosmides, & Tooby, 2016).

According to the ***signalling hypothesis***, third-party punishers would accrue indirect fitness benefits from their intervention via reputational gains. Third-party punishers' motivation to boost their reputation is indeed revealed by the so-called "audience effects". Specifically, individuals invest more resources in enacting 3PP when they are aware their decisions will be communicated to an audience (be it constituted by other participants or the experimenter) than when their decisions will remain anonymous (Kurzban, DeScioli, & O'Brien, 2007). Thus, 3PP would function as a mechanism to signal punishers' cooperative qualities, such as trustworthiness (Barclay, 2006; Jordan, Hoffman, Bloom, & Rand, 2016; Jordan & Rand, 2017), concern about group's shared values and social standing of the victim (Okimoto & Wenzel, 2009, 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, defined as enhanced ability to impose costs upon others (Lukaszewski, Simmons, Anderson, & Roney, 2016). Thus, 3PP might be akin to a strategy to assert dominance (Gordon, Madden, & Lea, 2014; Sylwester et al., 2013). If 3PP is used as a competitive tool, this would provide an explanation for why wrongdoers do not increase their investments in cooperation following punishment. Third-party punishers and their potential witnesses might interpret a change in the wrongdoers' behaviour as an act of subordination, which would lower the wrongdoers' social status within the group (Raihani & Bshary, 2019). To conclude, from an evolutionary perspective, in the case of cooperation signalling, third-party punishers would derive benefits from being appreciated by bystanders and thus preferentially selected in future interactions. Conversely, in the case of formidability signalling, punishers would benefit from being feared by observers, who will be consequently dissuaded from implementing any exploitive intentions they might have (Gordon et al., 2014; Raihani & Bshary, 2015a, 2015b).

Another account disputing group selection models of 3PP evolution is the aforementioned ***deterrence hypothesis***. Its proponents argue that 3PP might stem from a psychological mechanism evolved to prevent mistreatments from occurring to oneself or valued others (Delton & Krasnow, 2017). Since humans' ancestral social life was spent in small-scale societies characterised by high strength-of-ties and low mobility (Roos, Gelfand, Nau, & Carr, 2013), the capability to infer that mistreatment of one individual could predict subsequent personally-relevant mistreatment was likely to have adaptive value. In support of the deterrence hypothesis, it has been found that when no other information is available to the punishers, they base their punishment decisions on how they have seen third-party victims being treated by wrongdoers. When instead they have also access to information about how they would be personally treated by wrongdoers, this information overrides the one related to the treatment of third-parties in driving 3PP decisions (Krasnow et al., 2016).

Evidence that 3PP is also administered with the goal to deter misbehaviours from happening again to people the punisher has a welfare stake in comes from ethnographic (Boehm, 1987; Ericksen & Horton, 1992) and experimental studies (Lieberman & Linke, 2007; Pedersen et al., 2018). All these works are concordant in reporting that people preferentially act as third-party punishers on behalf of their family members or friends or at least in-group members, compared to strangers. Thus, in real-life settings and in experiments not laden with demands for punishment, interventions on behalf of strangers tend to be rare and practised when costs for punishers are low. When this type of intervention does occur, a plausible explanation is that third-parties' inaction would be negatively judged by an audience (Pedersen et al., 2018). In conclusion, if 3PP has evolved as a deterrent mechanism to protect oneself and important others from wrongdoers' misbehaviours, third-party punishers can reap both direct and inclusive fitness benefits from their interventions.

Whatever the resolution of this debate about 3PP's ultimate causes (i.e., adaptive functions), the present PhD thesis will be focused on its proximate causes (i.e., psychological mechanisms). However, a theoretical framework about 3PP behaviour would be incomplete if developmental investigations were fully disentangled from phylogenetic research (Tinbergen, 1963). The early onset of 3PP in children across different populations makes 3PP a good candidate for being considered a species-specific behaviour enabling human sociality. Nevertheless, such a definition would be premature without having verified whether 3PP is present in the behavioural repertoire of other animal species (Nielsen & Haun, 2016).

1.4. Comparative literature on third-party social evaluations and punishment

Whereas information about the ontogeny and evolution in our species of social norm enforcement, and 3PP in particular, has been rapidly accumulating over the past few years, research in the phylogenetic distribution of such behavioural reactions offers a still fragmented picture. Although this PhD thesis will not present new studies about the enforcement of putative social norms in non-human animals, the last section of this literature review will be dedicated to the comparative literature on the topic for completeness of information. As in the section dedicated to the developmental literature (section 1.2), here I will report comparative psychology findings comprising both mechanisms social norm enforcement is supposed to rely on: evaluation of third-party interactions and third-party punishment.

Regarding the experimental evidence of non-human animals' social evaluation skills, researchers have tested great apes – namely chimpanzees, bonobos (*Pan paniscus*), gorillas (*Gorilla gorilla*) and orangutans (*Pongo pygmaeus*) – by showing them interactions between actual humans or animated objects either in food-sharing or helping scenarios. It has been found that, even without being imparted any sort of training, chimpanzees – but not bonobos, gorillas and orangutans – spent significantly more time in proximity of a donor who had demonstrated generosity rather than selfishness towards a human begging for food (Russell, Call, & Dunbar, 2008). Conversely, in another food-sharing experimental paradigm, chimpanzees did not preferentially choose to beg from the generous over the selfish donor after having observed them interacting with a human or a chimpanzee beggar. Chimpanzees started exhibiting a preference for the generous donor in third-party contexts only after extensive training consisting in direct interactions with the two donors (Subiaul, Vonk, Okamoto-Barth, & Barth, 2008). In a following experiment, chimpanzees and orangutans – but not bonobos and gorillas – showed to approach more frequently a prosocial experimenter who had tried to give food to a human third-party compared to an antisocial experimenter who had prevented the food giving. However, the testing of great apes' third-party social evaluation skills was preceded by direct interactions between the great apes and the two experimenters (Herrmann, Keupp, Hare, Vaish, & Tomasello, 2013). A more recent study, modelled after the human infant paradigm developed by Hamlin et al. (2007), has analysed bonobos' third-party evaluations of animated objects in place of human adults. In the animations bonobos could see a recipient alternatively being helped by a prosocial agent and being hindered by an antisocial agent in reaching a goal. When required to reach for either the helper or the hinderer (preference for the neutral recipient was not assessed), surprisingly bonobos favoured the hinderer – a choice that may be indicative of attraction towards dominant individuals (Krupenye & Hare, 2018).

Among primates, studies about third-party social evaluations have also been carried out on monkeys, specifically tufted capuchins (*Sapajus paella*), common marmosets (*Callithrix jacchus*) and squirrel monkeys (*Saimiri sciureus*). Capuchins have been tested both in helping and reciprocity paradigms, whereas marmosets and squirrel monkeys only in reciprocity paradigms. Both experimental paradigms displayed interactions between humans, not between conspecifics or animated objects. In the helping paradigm a neutral person (i.e., recipient) was shown attempting in vain to open a container. Capuchin monkeys could either see the recipients' requests for help being responded to by a helper or ignored by a non-helper. When capuchins had to decide which of the two actors to accept a food offer from, they preferred the neutral recipient over the non-helper (negativity bias) but they did not exhibit any preference between the neutral recipient and the helper (lack of positivity bias). In a follow-up experiment to investigate whether capuchins are sensitive to actors' intentionality, the lack of a helping response was represented as either intentional (i.e., actor unwilling to help) or accidental (i.e., actor too focused on another task to notice the request for help). In this case, the subjects favoured the neutral recipient over the intentional non-helper, but did not manifest any preference between the neutral recipient and the accidental non-helper (Anderson, Kuroshima, Takimoto, & Fujita, 2013). In the reciprocity paradigm, instead, the capuchins watched a human actor (i.e., recipient) donating their resources to a second actor, who in turn decided to either reciprocate the act (i.e., reciprocator) or not (i.e., non-reciprocator). In the testing phase, the capuchins were required to choose which of the two actors to take food from. To note, differently from the helping paradigm, here the recipient is not a neutral person because they initiated a prosocial interaction with the other actor. It was found that capuchins preferred the recipient over the non-reciprocator (negativity bias), but were at chance when the choice was between the recipient and the reciprocator (lack of positivity bias). Interestingly, capuchins took the second actor's wealth into account when making their social evaluations. Indeed, a follow-up experiment revealed that the discrimination against the non-reciprocator occurred only when this actor actually refused the recipients' requests of reciprocation, not when they were incapable of fully reciprocating because of lack of initial resources (Anderson, Takimoto, Kuroshima, & Fujita, 2013). A similar reciprocity paradigm was also adopted with common marmosets and squirrel monkeys. Both species, like capuchins, exhibited a negativity bias by avoiding the non-reciprocator. Additionally, whilst marmosets did not have any positivity bias (like capuchins), squirrel monkeys surprisingly did also show a preference for the reciprocator over the recipient (Anderson, Bucher, Kuroshima, & Fujita, 2016; Kawai, Yasue, Banno, & Ichinohe, 2014).

Comparably to what Krupenye & Hare (2018) did with bonobos, McAuliffe at al. (2019) have recently adapted the human infant paradigm from Hamlin & Wynn (2007) for use with dogs. In this way it was possible to test whether dogs manifest a preference for non-human agents who behave prosocially versus antisocially in third-party virtual interactions. Specifically, McAuliffe at al. (2019) made dogs watch videos in which an animate object, while trying to reach a goal, was either assisted by a helper shape or impeded by a hinderer shape. Contrary to human infants and bonobos, dogs did not preferentially approach either the hinderer or the helper. However, they spent more time exploring the hinderer over the helper shape, possibly in an attempt to understand the reason for the agent's puzzling antisocial behaviour.

Another adaptation of Hamlin & Wynn's (2007) paradigm was produced by Johnson et al. (2018) to investigate bottlenose dolphins' (*Tursiops* spp.) social predictions in abstract contexts. In this experiment dolphins watched videos of a prosocial and an antisocial shape interacting with another shape, a neutral recipient. After the prosocial and antisocial shapes had moved off-screen in opposite directions, the recipient momentarily disappeared behind an occluder located at the centre of the screen. By analysing dolphins' anticipatory head turns as a proxy of their looking patterns, it has been demonstrated that dolphins expected the recipient to reappear from the occluder in correspondence with the side where the prosocial shape had last be seen. This indicated that dolphins predict that the recipient would follow and affiliate with a prosocial rather than with an antisocial agent.

Outside the class of mammals, third-party social evaluations have been investigated only in fish, specifically in the mutualism between the cleaner wrasse (*Labroides dimidiatus*) and its client reef fish. In the field, clients are more likely to invite inspections by cleaners they saw acting cooperatively towards other clients. Conversely, clients avoid cleaners they spotted cheating during the cleaning service provided to other clients (Bshary, 2002). In the laboratory, on the basis of the observation of third-party interactions, clients have been shown to spend more time in proximity to cleaners known to be cooperative compared to cleaners whose cooperative tendencies are unknown (Bshary & Grutter, 2006).

Coming to the comparative research about punishment, it has been shown that whereas 2PP is a widespread behaviour in the whole animal kingdom (Bshary & Grutter, 2002, 2005; Clutton-Brock, Russell, Sharpe, & Jordan, 2005; Clutton-Brock & Parker, 1995; Hauser, 1992; Jensen, 2010; Mulder & Langmore, 1993), actual 3PP seems to be a uniquely human characteristic (Riedl et al., 2012). However, behaviours resembling 3PP (which will be explained below) have been observed in eusocial insects like bees (Ratnieks & Visscher, 1989), wasps (Wenseleers et al., 2005) and ants (D'Ettorre, Heinze, & Ratnieks, 2004); cleaner fish (Raihani, Grutter, & Bshary, 2010; Raihani, Pinto, Grutter, Wismer, & Bshary, 2011); and non-human primates, both in apes (de Waal, 1982; Vervafcke, De Vries, & van Elsacker, 2000; Watts, 1997; Zucker, 1987) and monkeys (Kaplan, 1978; Kummer, 1967; Kurland, 1977; Petit & Thierry, 1994; Ren et al., 1991).

3PP-like behaviours in eusocial insects are often referred to as "worker policing" and consist in workers' interventions to defend the queen's reproductive primacy by destroying the eggs laid by other workers (Ratnieks & Wenseleers, 2005) or by attacking such workers (Ratnieks & Visscher, 1989). However, what could appear as an altruistic behaviour on the part of non-reproductive workers is in fact usually coerced by the colony's queen via chemical messages rather than being enacted voluntarily (Monnin, Ratnieks, Jones, & Beard, 2002; Ratnieks & Wenseleers, 2008), unlike actual 3PP in humans. Additionally, policing behaviour in eusocial insects bestows on the workers benefits in terms of inclusive fitness as they share more genes with the queen's offspring than with other workers' offspring (Monnin & Ratnieks, 2001).

The term "policing" is also used to describe third-party interventions carried out by non-human primates to interrupt conflicts arisen within their social group, usually over access to food or sexual partners (de Waal, 1982). Policing interventions are impartial because the arbitrator does not support either party involved in the conflict. In other words, policing in non-human primates, differently from both policing in eusocial insects and actual 3PP in humans, does not target specifically the transgressor. Performers of policing tend to be high-ranking individuals as they can afford not to fear redirected aggressions. On a proximate level, these individuals' third-party interventions might be motivated by a basic form of "community concern" (von Rohr et al., 2012). However, ultimate causes of impartial policing behaviour are most likely ascribable to the indirect fitness benefits the arbitrator gains from increased group stability (Flack, de Waal, & Krakauer, 2005; Flack, Girvan, de Waal, & Krakauer, 2006).

Another evidence of 3PP-like behaviour comes from the work on the mutualistic relationship between the bluestreak cleaner wrasses (*Labroides dimidiatus*) and their reef-fish clients: by removing ectoparasites from their clients' skin, cleaners obtain the majority of their nutrients. However, they occasionally “cheat” by eating off portions of their clients’ living tissues – which are preferred over ectoparasites as they have higher nutritional value. It has been observed that clients receive a higher quality cleaning service when they are inspected by male-female pairs of cleaners rather than by singletons (Bshary, Grutter, Willener, & Leimar, 2008). This is due to the fact that male cleaners punish their female partners when they catch them cheating. As a consequence, females increase their cooperation rates towards the clients in future joint inspections (Raihani et al., 2011). However, this behaviour does not constitute an example of actual 3PP because both the client and the male cleaner are victims of the female’s cheating act: every time it is bitten, the client swims away and the male cleaner thus loses the opportunity to eat ectoparasites. Therefore, by punishing the female, the male cleaner does not only benefit the client but he also derives direct fitness advantages in terms of food-provisioning from his investment in punishment (Raihani et al., 2010).

Experimental evidence of actual 3PP in non-human animals is scarce, even amongst our closest living relatives. To date, there have been only a handful of studies – all conducted in laboratory settings on captive chimpanzees (*Pan troglodytes*) – aimed at testing the extent of the sense of normativity in non-human animals by analysing the conditions triggering their punitive behaviours. In order to verify whether evolutionary precursors of 3PP behaviour are present in non-human animals, it is crucial to investigate whether they, like humans, intervene as uninvolved third-parties on behalf of victims suffering from a behavioural transgression. Von Rohr and colleagues (2011) have argued that an essential precondition for the evolution of social norms is the existence of what they call “personal norms”. Whereas *social norms* are defined by social expectations (empirical and/or normative) about the appropriate ways of acting towards anyone, *personal norms* are defined by personal expectations about the appropriate ways of acting towards the self. Below evidence of chimpanzees' behavioural reactions to violations of their personal and social expectations is reported.

When chimpanzees experience disadvantageous inequity as the result of a transgression committed at their personal expense – i.e., when a conspecific stole food from them – they show vengeful reactions in the form of 2PP against the transgressor. Conversely, when the disadvantageous inequity is not due to a transgression, such as in the case of viewing a conspecific eating desirable food they have never had access to, chimpanzees do not manifest any spiteful reaction against the other individual (Jensen, Call, & Tomasello, 2007). Moreover, after direct interaction with an antisocial agent who withdrew previously offered food, chimpanzees are willing to pay a cost (in terms of physical effort) to see the transgressor being punished by someone else, suggesting that the view of vicarious 2PP is for them a rewarding experience (Mendes, Steinbeis, Bueno-Guerra, Call, & Singer, 2018). Crucially, when dominant chimpanzees see a group member take food away from another individual, they do not intervene on behalf of the victim by enacting 3PP, not even when the victim is kin of theirs. This experimental evidence thus points at the possibility that chimpanzees lack sensitivity to behavioural transgressions affecting others than the self, at least in food-related contexts (Riedl et al., 2012).

Another study conducted on chimpanzees seems to suggest that these animals are instead capable of detecting behavioural transgressions affecting others than the self or in-group members, but in harm-related contexts. However, this capability to detect harm transgressions is not usually accompanied by increased emotional reactions towards the victims. Specifically, in this study chimpanzees' looking patterns and physiological arousal were measured while they were shown a video containing scenes of an infanticide committed by conspecifics unknown to them. It was found that chimpanzees looked longer at infanticide scenes compared to control scenes displaying hunting, nut cracking or aggressive interactions among adults. But crucially this increased attention did not translate into increased arousal, possibly because the transgression affected out-group rather than in-group members (von Rohr, van Schaik, Kissling, & Burkart, 2015). Therefore, chimpanzees might have empirical but not normative expectations about out-group members' behaviour. It remains to be verified whether chimpanzees possess normative expectations about in-group members' behaviour in harm-related contexts.

Of note, absence of evidence of 3PP and normative expectations in non-human animals is not necessarily evidence of their absence. For example, chimpanzees might have normative expectations but, at the same time, it might not be possible for them to express such expectations through 3PP because of social factors. Specifically, when not in a position of power, bystanders to norm violations might fear retaliation and redirected aggressions upon their potential intervention (von Rohr et al., 2011). However, this does not explain why, when a conflict arises, dominant individuals have been observed enacting impartial policing behaviour but not 3PP expressly directed at the transgressor.

In conclusion, amongst non-human animals there is remarkable cross-species variability in the ability to evaluate, from a third-party perspective, other individuals on the basis of their adherence to putative norms of behaviour (Abdai & Miklósi, 2016). When it comes instead to reacting to behavioural transgressions, non-human animals are prompt in inflicting punishment upon transgressors when they themselves are the victims, but not when they merely are unaffected bystanders (Riedl et al., 2012). It has been demonstrated that non-human animals' behaviours that have been previously mistaken for 3PP, in reality, bring about direct (rather than indirect) fitness benefits to the intervener, or do not specifically target the transgressor, or are enacted under coercion (Jensen, 2010). Therefore, although further research in comparative psychology is surely needed, actual 3PP most likely is a uniquely human behaviour whose function is still under debate (as seen in section 1.3). This implies that, even though non-human animals have clear expectations about how they want to be treated (Jensen et al., 2007; Riedl et al., 2012) and can even predict how others prefer to be treated (Johnson et al., 2018), they do not experience any sense of obligation (or “ought” feeling) to correct or discourage behavioural transgressions negatively affecting other individuals’ lives (Burkart, Bruegger, & van Schaik, 2018; de Waal, 2014).

1.5. Overview of the thesis

Having introduced elements of comparative and evolutionary psychology to appropriately contextualise 3PP behaviour, I now focus again on developmental research and address the gaps in its literature where further contribution is intended here.

From the review of children’s 3PP literature presented at the beginning of this Chapter, several issues had emerged as un- or under-investigated. Firstly, the range of norm violations adopted to elicit 3PP has been limited to injunctive (moral) norms, and specifically issues of fairness and harm. Since there is no guarantee these results are unaffected by descriptive norm violations and generalisable to other moral domains, an important advance would be to understand whether 3PP is influenced by the kind of norm violations punishers witness. Secondly, moral violations have so far been operationalised as negative outcomes for the victim, mostly stemming from the transgressor’s negative intentions. Only one study (Gummerum & Chu, 2014) examined whether children are capable of integrating intention and outcome information in their 3PP decisions. However, this study was conducted by adopting the “strategy method”, which requires participants to make punishment decisions before they witness a transgression. It is yet to be verified whether the integration between intentions and outcomes would be facilitated by the use of more naturalistic methods, in which 3PP is enacted after third-parties have viewed a moral transgression. Moreover, whereas the role of emotions as antecedents of 3PP has begun to be clarified (Gummerum, López-Pérez, et al., 2019), the emotional consequences of 3PP constitute a completely unexplored topic in developmental research: it needs to be understood whether children derive, or at least expect to derive, satisfaction from meting out 3PP against transgressors. Hedonic experiences or expectations of 3PP would be in line with a retributive motivation to punish (Crockett et al., 2014), but less likely to be consistent with a deterrent motivation. Another potential 3PP motivation, in addition to retribution and deterrence, is represented by reputation enhancement. Thus, the way to elucidate the role of reputation is through establishing whether children are responsive to the perceived presence of an audience when they make their 3PP decisions.

In order to deepen the knowledge about children's 3PP behaviour, I thus designed a series of four experiments, differing in the specific research questions and experimental paradigms adopted, as well as on the basis of the age range, country of residence and socio-economic status of the sample.

Experiments 1-3 were conducted on a wide age range, including 5- to 11-year-old children, who acted as uninvolved referees in a computer game named *MegaAttack*. The game was developed from scratch by the research team and featured teams of internet players supposedly playing live. As these players violated fairness or loyalty norms, respectively an individualising and a binding moral domain (Graham et al., 2013; Graham, Haidt, & Nosek, 2009), children were offered the opportunity to punish them. Experiment 4 was instead conducted on a slightly narrower age range, ranging from 7- to 11-year-old children, who operated a Justice System in which they viewed moral transgressions of various types (physical harm; property destruction; theft; disloyalty/inequity; sanctity/authority transgression; deception/liberty violation) in *Minecraft*, a globally popular videogame. As impartial judges of the Justice System, children could respond to transgressions with two types of third-party interventions, namely punishment of transgressors and compensation of victims. Regarding the demographics of the samples, children participating in Experiments 1-2 lived in the UK and on average came from a highly educated middle-class background. Notably, the UK is considered a WEIRD country, whereby "WEIRD" is a cultural identifier standing for "Western, Educated, Industrialised, Rich and Democratic" (Henrich, Heine, & Norenzayan, 2010). The sample of Experiment 3 was instead constituted only by non-WEIRD children, specifically urban Colombian children of generally low socio-economic status. Finally, Experiment 4 was carried out cross-culturally, recruiting children from the UK, Colombia and Italy of various social backgrounds. It is notable that Italian culture is ranked somewhere between British individualistic and Colombian collectivistic culture on a number of validated scales (Hofstede, 2001; House, Hanges, Javidan, Dorfman, & Gupta, 2004).

Below are reported the research questions generated from the reviewed literature that I decided to investigate in the present PhD thesis, subdivided by chapters and experiments.

Chapter 2 includes Experiments 1 and 2. Both of them were aimed at examining the developmental trajectory of children's 3PP severity and clarifying whether children make the type of 3PP (e.g., economic or social punishment) fit the type of moral transgression (e.g., violation of fairness or loyalty norms) in terms of moral domains. Experiment 1 was additionally designed to verify whether deviations from descriptive norms (i.e., what is commonly done) along with deviations from moral norms (i.e., what should be done) cause harsher judgements of transgression severity and 3PP decisions. Experiment 2 was also intended to explore the affective states children experience in enacting 3PP, as well as to evaluate audience effects on children's judgements of transgression severity and 3PP severity by presenting them with cues of observability.

In **Chapter 3** I introduce Experiment 3. Differently from Experiments 1-2, in Experiment 3 the developmental trajectory of children's 3PP severity was analysed across moral domains (e.g., unfairness and disloyalty). Audience effects on children's judgements of transgression severity and 3PP severity were evaluated by presenting children with cues of accountability along with cues of observability. Experiment 3 also addressed whether children's 3PP-related affective states are influenced by time (i.e., before, during and after punishment allocation) and task wording (i.e., emphasis vs lack of emphasis on punishment outcomes for the transgressor). Moreover, an entirely new issue was investigated in Experiment 3 in comparison to Experiments 1-2, namely the integration between outcome and intention information in children's judgements of transgression severity and 3PP severity along the developmental trajectory and across moral domains.

In **Chapter 4** I present Experiment 4, which was intended to clarify the extent to which children's compensatory and punitive decisions (i.e., 3PP severity, compensation level, 3PP vs compensation endorsement) are determined by judgement of transgression severity, type of moral transgression and children's age. Experiment 4 was also designed to examine whether children's punishment- and compensation-related enjoyment are affected by a variety of factors, such as: judgement of transgression severity, type of third-party intervention, time passed since the intervention, children's age and retribution vs deterrence endorsement. Moreover, in order to investigate children's 3PP explicit justifications, it was assessed whether retribution vs deterrence endorsement is influenced by framing messages (i.e., emphasis on retribution or deterrence or compensation), children's age and country of residence. Finally, in order to shed light on children's 3PP implicit motivations, Experiment 4 was also aimed at evaluating the effects of framing messages on children's compensatory and punitive decisions.

The aforementioned three experimental chapters – Chapter 2, 3 and 4 – will be roughly structured as papers, including an introduction with the literature related to each specific research topic, a methodology section with the details of each experimental paradigm, followed by the presentation and discussion of the results. The final chapter of the present PhD thesis – **Chapter 5** – will bring together the findings of the four experiments in order to draw the conclusions deriving from this research effort and identify the avenues for future investigations.

Chapter 2: Experiments 1-2

A SENSE OF MORAL DUTY AND INEQUALITY AVERSION MOTIVATE CHILDREN'S THIRD-PARTY PUNISHMENT

Experiments 1-2 combined

2.1. General introduction

In light of the literature review on children's 3PP behaviour (see Chapter 1), Chapter 2 will present two experiments that were designed to investigate the following issues: whether children tend to fit the kind of punishment to the kind of moral transgression in terms of moral domain (Experiments 1-2); whether they punish violations of descriptive norms (what is commonly done) as well as moral violations (Eriksson, Strimling, & Coultas, 2015) (Experiment 1); whether their 3PP responses 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. To my knowledge, mine is the second line of research employing a virtual game to investigate children's social cognition (after the virtual ball-toss game *Cyberball* to study ostracism in children; Crowley, Wu, Molfese, & Mayes, 2010).

Experiment 1

2.2. Introduction

2.2.1. Social norm typology: moral domains, and descriptive vs injunctive norms

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; Gray, Young, & Waytz, 2012; 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" and "Morality-as-Cooperation" 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., 2009). Morality-as-Cooperation theory is based on a seven-factor model of morality: *family*, *group loyalty*, *reciprocity*, *bravery*, *respect*, *fairness* and *property* (Curry, Chesters, & Van Lissa, 2019). 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. I propose two rival cognitive models: the deep separation model and the domain general model. According to the *deep separation model*, transgressions of different domains lead to different types of punishment motivation, potentially motivating different types of punishment behaviour (Figure 1A). According to the *domain general model*, 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, I investigated whether children tend to react differently to different types of moral norm violations. I 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 deep separation model to the test, I hypothesised 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. I 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). Thus, *moral norms* could be considered a subset of injunctive norms. People tend to make inferences from descriptive norms to injunctive norms and vice versa, falling victim to the so-called "naturalistic fallacy" (Hume, 1739/2000). In that respect, it has been found that when an injunctive norm transgression is described as common, people express more lenient judgements. Conversely, when the same transgression is described as uncommon, people's judgements are harsher (Eriksson et al., 2015; McGraw, 1985; Trafimow, Reeder, & Bilsing, 2001; Welch et al., 2005).

From a developmental perspective, this descriptive-to-injunctive tendency has also been observed in young children involved in behavioural experiments. For example, after being shown how a demonstrator operates an apparatus (i.e., descriptive norm), preschool children employ injunctive language (''You *should* do this'') to protest against those who do not faithfully imitate the demonstration (Kenward, 2012). Interestingly, children can make these descriptive-to-injunctive inferences even after only one exposure to the model's action (Schmidt, Butler, Heinz, & Tomasello, 2016). Furthermore, recent studies have demonstrated that young children negatively evaluate individuals who do not conform to their group-descriptive norm (e.g., type of food eaten), and that such negative judgements are justified by injunctive norm-based explanations (Roberts, Gelman, & Ho, 2017; Roberts, Guo, Ho, & Gelman, 2018; Roberts, Ho, & Gelman, 2017).

It remains to be understood whether children's negative evaluations of descriptive norm violations can translate into 3PP towards individuals who do not conform to their group-descriptive norms, but on the basis of the above literature I predict that it may do so. However, violations of purely descriptive norms might cause weak punitive motivations, below the threshold for action. Therefore, rather than investigating whether a descriptive norm violation alone motivates punishment, I investigated whether descriptive norm violations would increase the severity of punishment allocated for moral norm violations. Because substantial variance in punishment severity is typically explained by judgements of transgression severity (Alter, Kernochan, & Darley, 2007), I measured and controlled for transgression severity judgements when modelling punishment severity.

2.2.2. Age effect on third-party punishment

As highlighted in the developmental literature presented in Chapter 1 (section 1.2), 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 Melanesian children (age range: 4 to 10 years of age) who witnessed antisocial actions, i.e. destruction of a gift (Dixson & Kenward, *in prep*), as well as in Western children who watched unfair allocations made during a Triadic Dictator Game. This latter 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). By contrast, the Triadic Dictator Game study conducted by Gummerum, Takezawa & Keller (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, I predicted I would detect the same in Experiment 1 even though I measured children's punitiveness in terms of 3PP severity instead of 3PP rates.

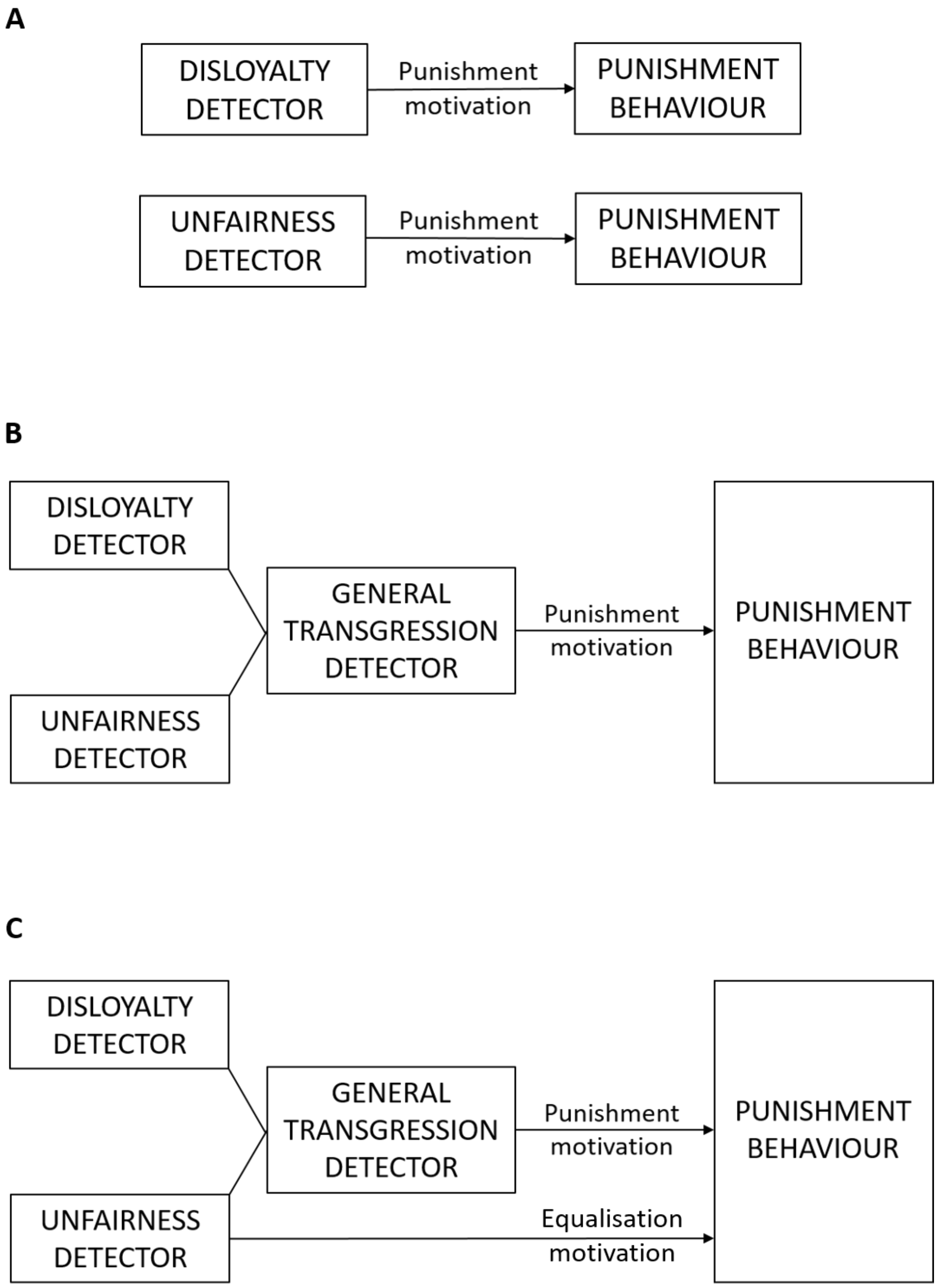

***Figure 1*. Cognitive models of punishment motivations illustrating the relationship between transgressions in different moral domains and consequent punitive outcomes.** A) Deep separation model. B) Domain general model. C) Domain general plus equalisation model.

2.3. Method

2.3.1. 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).

2.3.2. Sample

Participants were 72 primary school-aged children (*mean age*: 8.83 ± 1.81 years; *age range*: from 5.45 years to 11.95 years; 32 females and 40 males). Six additional children were tested but had to be excluded from the sample: three for exceeding the set age limit (i.e., not older than 11 years); one for a technical failure in the equipment; one for unwillingness to continue; and one for difficulties in understanding the experimenter's requests. They were 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). The stopping rule was set roughly half-way through data collection: one cell in the eight-cell table for counterbalancing of irrelevant variables had reached 9 participants, so it was decided to continue until all cells had 9 participants. The study was approved by the Oxford Brookes University Ethical Review Committee under the project name "Children's social judgement in a computer game" (Study Number 171101).

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 ethnicity (*parents' nationality*: 23 British, 10 non-British, 2 unspecified).

2.3.3. Design

I adopted a 2x2 fully within-subject design in which the factors were *Descriptive norm violation* (descriptive norm conformity vs. descriptive norm violation) and *Type of moral transgression* (fairness vs. loyalty violation). I 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 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 Appendix A – Table α1 for details). Each test-trial featured a different pair of player avatars (different animals inside space-ships).

The dependent variables measured were: *type of punishment*: 2 levels (economic, loss of gems as an in-game resource vs. social, banning from the game); *severity of punishment* (6 ordinal levels; for social punishment: 0 minutes, 1 minute, 5 minutes, 20 minutes, 1 hour, 1 day; for economic punishment: 0, 2, 5, 10, 50, or 100 gems); *judgement of transgression severity*: 5 ordinal levels (from 1 = "just a little bad" to 5 = "super bad").

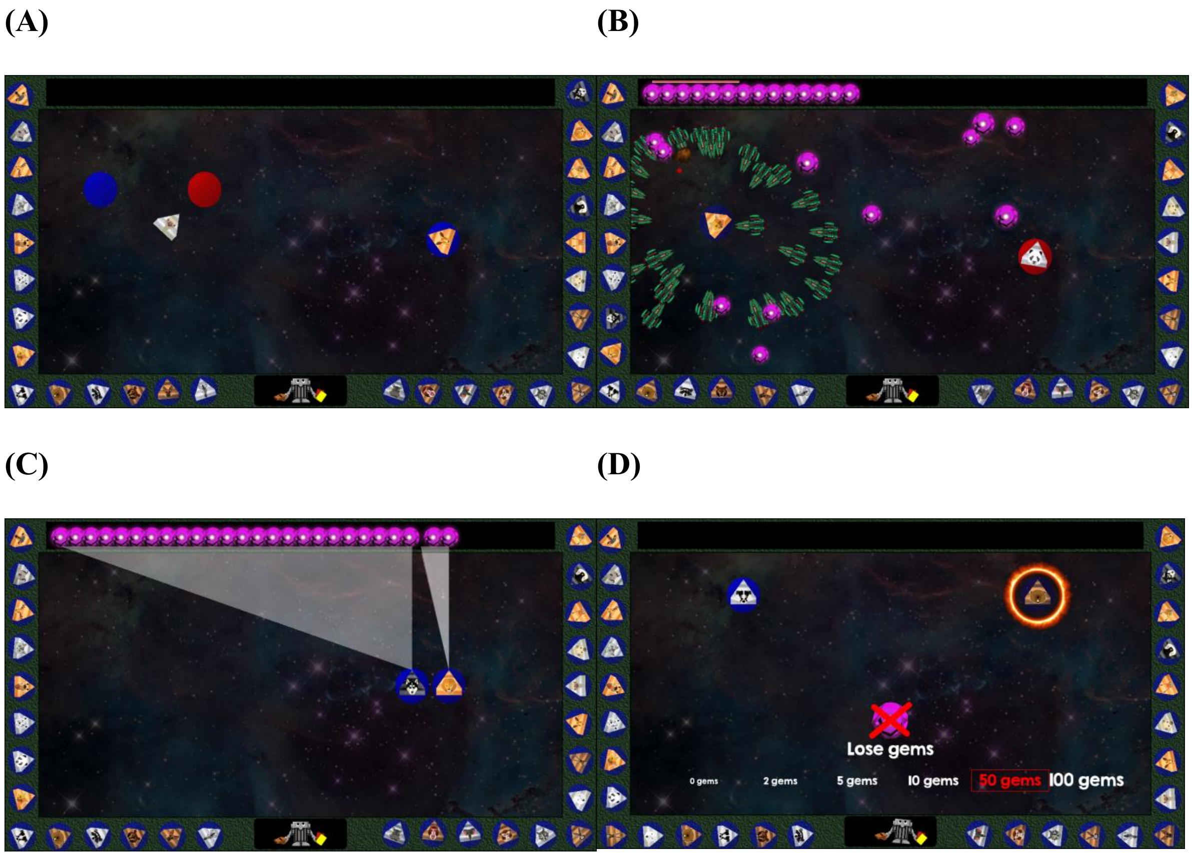

Figure 2. **Different stages of Experiment 1 game bouts.** (A) Shield-choice stage: player Ostrich makes a descriptively non-normative choice. (B) Gem-collection stage: player Fox is under deadly threat from a Mega-attack, as disloyal player Panda ignores the situation and continues to collect gems. (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.

2.3.4. Procedure

The procedure was divided into three phases (see full script in Appendix A – section α1 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 section 2.2.3 about 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, 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). Descriptive norm-violations happened when one of the players chose for themselves an uncommon shield colour (Figure 2A).

After each of the five internet scenarios shown (**1 refereeing familiarisation + 4 test trials**), in a refereeing stage the child answered for each of the two players in turn: "*Did they do anything wrong?*" If misbehaviour was identified, the child had to judge the severity of the norm-transgression using a 5-point smiley face scale (ranging from "super bad" to "just a little bad"), as well as to decide the type (social or economic) and severity of the punishment (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.

2.4. Results & Discussion

All analyses were conducted in the R programming language (RStudio Team, 2015).

2.4.1. Preliminary analyses

2.4.1.1. 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 (see Appendix A – section α4 for supplementary statistical analyses of Experiment 1). Therefore, for the statistical analyses data is included irrespective of believability.

2.4.1.2. 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.

2.4.2. Main analyses

2.4.2.1. Choice of punishment types

I 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 sign-tests were conducted (excluding .5 values which are uninformative in this one-sample context). In unfairness trials values of 1 occurred significantly more often (n = 39) than values of 0 (n = 11), $p < .001$, sign-test, whereas in disloyalty trials the difference between the occurrence of values of 1 (n = 30) and values of 0 (n = 17) approached but did not reach significance, $p = .079$.

In order to investigate the effects of age on the tendency to make the punishment fit the crime, I also calculated an overall "Punishment Fits The Crime" (PFTC) score, as the mean of the above two proportions for each individual. This score did not change as a function of age, $F(1,70) = 1.05$, $p = .308$, $R^2 = .01$, in contrast with my 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 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, ruling out the pure *domain general model*, 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 *deep separation model*, according to which specific transgressions motivate specific punishments across domains (Figure 1A), is therefore also lacking. Different interpretations can explain the detected effect. 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 model*). Alternatively, children's punishment behaviour might have been additionally motivated by inequality aversion, with economic costs imposed not only to punish but also to correct unjust resource distributions (*domain general plus equalisation model*; Figure 1C). 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 model and the domain general plus equalisation model because they both predict 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).

2.4.2.2. Effects of descriptive norm violations

Judgements of transgression severity were not greater when the descriptive norm was violated (M = 3.53, SD = 0.99) than when the descriptive norm was adhered to (M = 3.49, SD

= 0.98), t(71) = 0.32, *p* = .747, d = 0.04, 95% CI for d [-0.19, 0.27]. Neither was punishment severity greater when the descriptive norm was violated (M = 4.39, SD = 0.99) than when the descriptive norm was adhered to (M = 4.33, SD = 1.00), t(71) = 0.62, *p* = .537, d = 0.07, 95% CI for d [-0.16, 0.30]. These null results could seem surprising in view of the previous literature that found that children evaluate individuals who do not conform to group-descriptive norms more negatively than those who do conform (Roberts, Gelman, et al., 2017; Roberts et al., 2018; Roberts, Ho, et al., 2017). However, previous research did not present children with deviations from descriptive norms along with deviations from injunctive norms as it was done in Experiment 1. What can be concluded is that transgressions resulting from violating *both* a descriptive and injunctive norm were not judged more severely than *purely* injunctive transgressions.

To note, since it was important not to interrupt the flow of the experiment for believability purposes, no manipulation check at the end of each test trial was included to verify whether children had detected the descriptive violations. Nevertheless, I am confident that the null effect of descriptive violations on judgements of transgressions severity and on punishment severity was not due to children not registering the descriptive violations, as players' uncommon shield choices were explicitly pointed out by the voice-over.

2.4.2.3. Severity in punishment across development

A multiple linear regression was conducted to predict punishment severity based on children's age in years, controlling for their judgement of transgression severity, F(2,69) = 9.50, *p* < .001, R^2 = .22. Punishment severity was negatively predicted by age in years (B = -0.15, *p* = .009) and positively predicted by judgement of transgression severity (B = 0.32, *p* = .006). Thus, it was demonstrated that this decrease in children's severity in punishment with increasing age was not due to older children considering the transgressions less serious.

This result was at odds with previous research analysing punishment rates across development, and is discussed after replication in Experiment 2.

Experiment 2

2.5. Introduction

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

2.5.1. 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 domain general model in which equalisation motives also influence behaviour (Table 1). 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 account predicts that the economic punishment of a gem fine would now be associated with loyalty rather than fairness transgressions. The unfairness now concerned a different resource (bombs) that could no longer be equalised by a gem fine. The domain general plus equalisation model therefore predicts no preference for either type of punishment in either condition (Table 1).

***Table 1.* Predicted punishment preference results for each condition according to different models, plus observed results.**

Condition	**Deep separation**	**Domain general**	**Domain general plus equalisation**	**Associative**	**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)	Detection of violation of any domain motivates general punishment behaviour but equalisation motives can modify behaviour (Fig. 1C)	Salient element of transgression primes punishment involving same element	
Exp. 1 Loyalty transgression	Social punishment	No punitive preference	No punitive preference	No punitive preference	No punitive preference (trend, p < .1, toward social punishment)
Exp. 1 Fairness transgression	Economic punishment	No punitive preference	Economic punishment[a]	Economic punishment[b]	Economic punishment
Exp. 2 Loyalty transgression	Social punishment	No punitive preference	No punitive preference	Economic punishment[c]	No punitive preference (trend, p < .1, towards economic punishment)
Exp. 2 Fairness 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).

2.5.2. Audience effects on moral behaviour and judgements

Audience effects – namely, behavioural changes induced by the presence of an audience or cues of observation – have been extensively investigated in adults but to a lesser extent in children. Adults who feel they are being watched modify their own behaviour in order to meet their audience's expectations. For instance, people are more likely to engage in moralistic punishment when in the presence of an audience (Kurzban et al., 2007; Piazza & Bering, 2008). Children have also shown increased moral behaviour in the presence of an audience, real or believed as such. Five-year-olds are more generous when recipients are visible (Leimgruber, Shaw, Santos, & Olson, 2012). Children as young as age 5 are also more likely to share and less likely to steal (Engelmann, Herrmann, & Tomasello, 2012), cheat (Piazza, Bering, & Ingram, 2011) and lie (Fu, Evans, Xu, & Lee, 2012) when someone is watching than when they are alone. Children as young as 6 are less likely to be fair to others when they can do so without appearing to be unfair (Shaw et al., 2014).

When no actual audience is present but only implicit cues of observation – such as images of "watching eyes" – adults seemed to unconsciously regulate their behaviour (e.g., Haley & Fessler, 2005) and moral judgements (Bourrat, Baumard, & McKay, 2011), although this line of research has produced mixed results (Dear, Dutton, & Fox, 2019; Northover, Pedersen, Cohen, & Andrews, 2017; Pfattheicher & Keller, 2015). Null results have partly been explained in terms of exposure length to the observation cue (Sparks & Barclay, 2013) and anonymity (i.e., people can ignore observation cues if they are not held personally accountable for their actions, Raihani & Bshary, 2012). Regarding, instead, the effect of cues of observation on children's moral behaviour, there is indication that the way in which "watching eyes" are presented (i.e., how explicit the exposure is) is the determining factor for finding or not finding an audience effect. Indeed, images of eyes implicitly presented were not effective in increasing children's generosity in a one-shot dictator game

(Fujii, Takagishi, Koizumi, & Okada, 2015; Vogt, Efferson, Berger, & Fehr, 2015). However, Kelsey, Grossmann and Vaish (2018) observed a significant increase in children's sharing tendency when the exposure to "watching eyes" switched from implicit to explicit.

Thus, the potential effects of observation cues on a broader range of children's moral behaviours, including 3PP, are of interest. For this reason in Experiment 2, once the child took the role of the referee, it was manipulated whether or not children were subject to cues of being watched. Because many potential mechanisms are involved, from strategic cognition concerning the expectation of approval to more automatic tendencies, in this first investigation of the topic no attempts to distinguish different types of mechanism were made. Rather, a collection of audience cues were manipulated together – 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 third-party punishment against norm violators, and express more severe judgments about transgressions, in the Audience condition.

2.5.3. 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, López-Pérez, Van Dijk, & Van Dillen, 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 my 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. Punishers are often conflicted with complex emotions that may also vary in quality and intensity across time.

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 punish the wrongdoer. Further, punishers who saw the wrongdoer evidence contrition in response to punishment experienced an increase in satisfaction (Funk et al., 2014; Gollwitzer, Meder, & Schmitt, 2011).

Another line of research has examined the relative contribution of retribution and deterrence motivations in 2PP and 3PP. 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, Özdemir and Fehr (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. It turned out that participants in the hidden punishment condition sanctioned the wrongdoer almost as frequently as in the open punishment condition, both in 2PP and 3PP contexts. This showed that 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), I 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, I 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, I 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.

2.6. Method

2.6.1. Sample

Participants were 80 primary school-aged children (*mean age*: 7.91 ± 1.62 years; *age range*: from 5.27 years to 11.56 years; 23 females and 57 males); one additional child was tested but had to be excluded for unwillingness to continue. Children were 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. The stopping rule was to collect as much data as possible by a given end date.

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).

2.6.2. Design

I 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).

I 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 presence or absence (see Appendix A – Table α2).

The dependent variables measured were: *type of punishment* (gem fine or a ban as in Experiment 1, except the opportunity to choose no punishment was now also present); *severity of punishment* (6 ordinal levels as in Experiment 1); *affective state in enacting punishment* (11 ordinal levels from -5, "very bad", to +5, "very good"); *judgement of transgression severity* (using the same 11-point scale, differing from Experiment 1 in order to use only one valence scale in this experiment and thus avoid confusion for the children).

2.6.3. 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 for example 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.

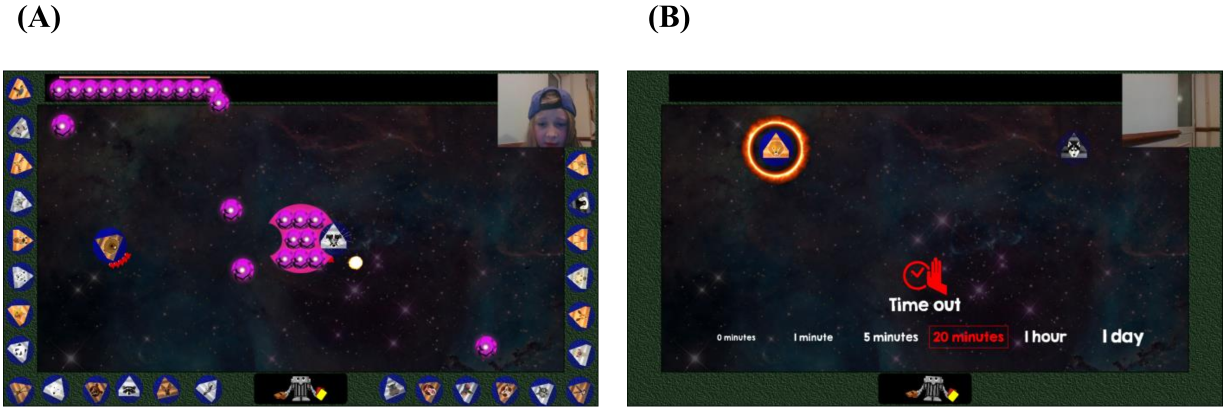

***Figure 3.* Experiment 2 game bouts stages with differences to Experiment 1.** (A) Gem-collection stage: player Badger is stuck on the Mega-gem and taking damage from enemies, as disloyal player Beaver refuses to release them by also attaching to the Mega-gem to collect the gems, and the thumbnailed live-streamer observes and commentates. (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.

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 3A). 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 (see Figure 3A). 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 Appendix A – section α3.5 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 conditions 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 conditions 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 3B). 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, I 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.

2.7. Results & Discussion

2.7.1. Preliminary analyses

2.7.1.1. 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 (see Appendix A – section α5 for supplementary statistical analyses of Experiment 2), therefore for the statistical analyses data is included regardless of believability.

2.7.1.2. 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.

2.7.1.3. 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 considered in the analyses.

2.7.2. Main analyses

2.7.2.1. Audience effects

Judgements of transgression severity in the Audience condition (M = -3.23; SD = 1.17) were significantly harsher than those expressed in the No Audience condition (M = -3.00; SD = 1.14), t(78) = -2.25, *p* = .027, d = -0.25, 95% CI for d [-0.47, -0.03]. This result is thus in accordance with Bourrat et al.'s (2011) study, which had found significant differences between adults exposed to eye images and controls in the degree of severity of their judgements towards moral transgression vignettes.

However, children's punishment severity in the Audience condition (M = 4.49; SD = 1.20) was not significantly higher than in the No Audience condition (M = 4.36; SD = 1.33), t(74) = 1.10, *p* = .274, d = 0.13, 95% CI for d [-0.09, 0.35]. This null result is in contrast with findings of Kurzban et al. (2007) and Piazza and Bering (2008), who observed an increase in moralistic punishment when adult participants thought their reputation was at stake. It is plausible that Experiment 2's audience manipulation, whilst including cues of observability from three different sources – the live-streamer, the audience of avatars and the experimenter – was insufficient to elicit children's reputation management because they nevertheless felt anonymous. Two of three manipulation cues were internet-based, implying participants may not have felt personally identifiable as making moral choices.

No manipulation check after the test trials was included because it would have interfered with believability. Nonetheless, I am confident that children registered at least one cue: the live-streamer's presence (and disappearance in the No Audience condition for the punishment stage) were highly salient. I therefore doubt that the absence of a detected audience effect was due to total failure of the manipulation. I consider it more likely that a virtual environment in which children's anonymity is protected has the potential to nullify the effectiveness of an audience in modifying children's 3PP and moral judgements. It is open to future investigations whether an audience manipulation making children feel not only observed but also personally accountable would affect their severity in punishing and/or their judgements of transgression severity.

2.7.2.2. Affective states

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 4). 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% among children who believed they were really punishing, 58% among children who believed they were warning players about future punishment, and 50% 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.

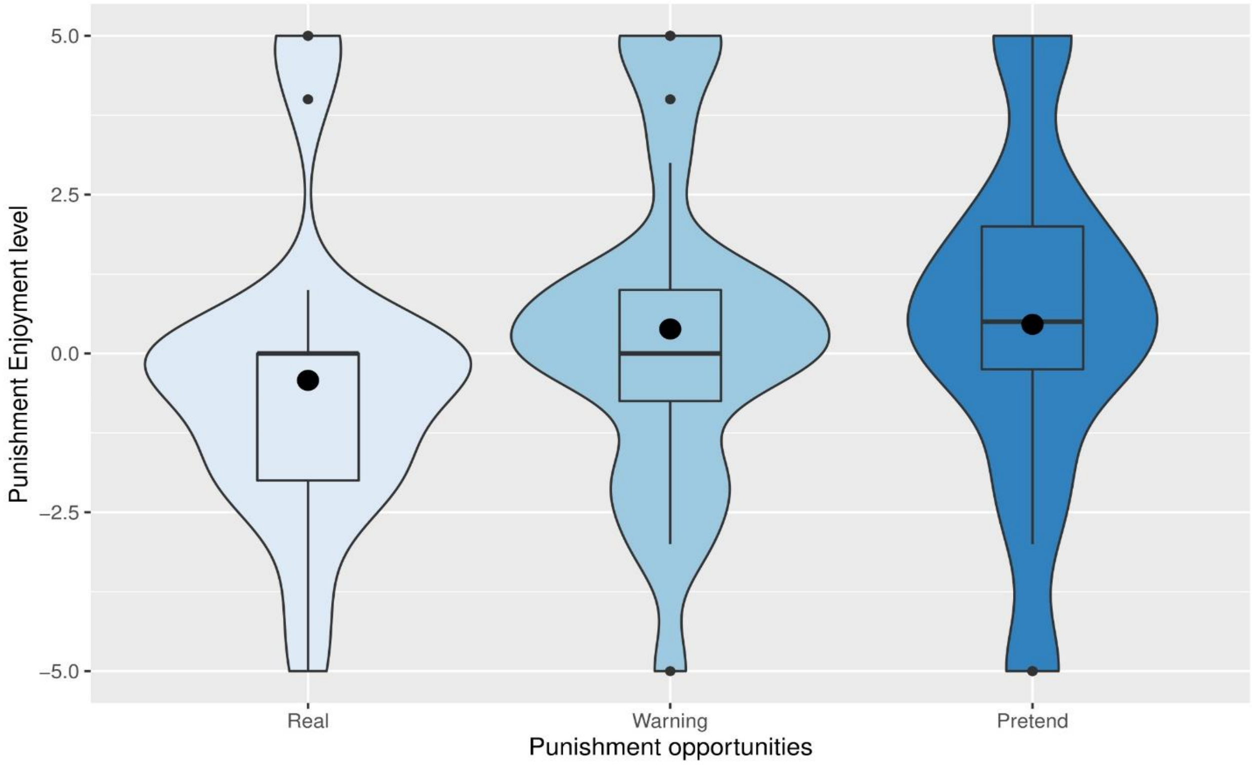

***Figure 4.* Experiment 2 punishment enjoyment by punishment opportunity condition: real; warning; pretend.** Violin plots wrapping boxplots; boxplots showing median and interquartile range, outliers, and a large dot for mean value.

This result accords with Carlsmith et al.'s (2008) finding that punishing potentially has a negative impact on affective states, extending this result from adults to children: in their experiment punishers experienced more negative affective states than non-punishers. This result was particularly surprising in the light of Hao et al.'s (2016) finding that 3PP is associated with positive affect in adolescents. This lack of enjoyment of punishing 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 punishment 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.

2.7.2.3. 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. In unfairness trials values of 1 and values of 0 were equally occurring (n = 22), indicating no tendency for the punishment domain to match the transgression, while in disloyalty trials the difference between the occurrence of values of 1 (n = 16) and values of 0 (n = 29) indicated a non-significant trend for economic punishment (gem fine) to be chosen for the social transgression (betrayal at the mega-gem), $p = .072$, sign test.

Thus, while the results of Experiment 1 were predicted by both the *associative* and *domain general plus equalisation models*, the direction of the non-significant trend for gem-related disloyalty to be punished by a gem fine in Experiment 2 was predicted only by the *associative model*. The combined results of Experiments 1 and 2 render the associative model very plausible, according to which the preference is for punishment that is connected to salient but superficial features of the transgression. However, it is not possible rule out the *domain general plus equalisation model* as this is also consistent with the Experiment 1 results and not ruled out by Experiment 2. Therefore, although the *deep separation model* can be considered discarded (because of inadequate evidence in Experiment 1 and tentative contrary evidence in Experiment 2), it is not possible to accurately distinguish between associative and equalisation explanations. Further research is needed to shed light on this issue.

Finally, in order to investigate the effects of age on the tendency to make the punishment fit the crime, I also calculated an overall "Punishment Fits The Crime" (PFTC) score, as the mean between the proportion of unfairness trials sanctioned with economic punishment and the proportion of disloyalty trials sanctioned with social punishment. This score did not change as a function of age F(1,75) = 0.00, p = .966, R^2 < .001, confirming the result of Experiment 1.

2.7.2.4. Severity in punishment across development

In order to verify whether the developmental trajectory of children's severity in punishment in Experiment 2 was similar to that observed in Experiment 1, I ran a multiple linear regression to predict this variable based on children's age, controlling for children's judgement of transgression severity. A significant regression was found, F(2,73) = 13.46, p < .001, R^2 = .27. Again, children's severity in punishment decreased with age in years (B = -0.28, p < .001), and was significantly related to judgement of transgression severity (B = -0.29, p = .014).

The finding that punishment severity is negatively predicted by age was somewhat unexpected, considering that the majority of previous literature has focussed on children's 3PP rates (i.e., probability to engage vs not engage in punishment) instead of 3PP severity, demonstrating that they increase rather than decrease with age (Dixson & Kenward, *in prep;* Jordan et al., 2014; McAuliffe et al., 2015; Salali et al., 2015). It is therefore 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 replicated finding that punishment 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. suggest decreases in punishment severity are linked to emotional development, and in line with this I suggest that the observed decrease with age of punishment 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 (measured through fMRI scanner) 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 punishment behaviour. 2.

Experiments 1-2 combined

2.8. General discussion

My investigation has shed light on children's 3PP by making use of an innovative computerised paradigm that simplified the manipulation of numerous variables. In this way, I tested cognitive models of punishment motivations, and examined potential mediators of 3PP such as descriptive-to-injunctive inferences, affective reactions, age and audience presence.

I advanced knowledge about children's punitive responses to moral violations in different moral domains in several ways. Firstly, I established that children punish loyalty violations similarly to fairness violations. Secondly, Experiments 1-2 provided evidence suggesting that there is no deep 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. Thirdly, I found that although the basic motive to punish is therefore moral-domain-general, punishment behaviour can be modified by salient aspects of the transgression. Together, the results provide some evidence for two different processes that may be responsible for such modification. That matching of the punishment to the crime was unambiguous only when the punishment could mitigate the crime (Experiment 1, gem fine for gem unfairness) is consistent with children's well known equalisation concerns (Gummerum & Chu, 2014; Gummerum, López-Pérez, et al., 2019; Jordan et al., 2014). However, this result is also consistent with an associative account: children simply match punishment to crime in terms of the objects involved. Only this account predicts the near significant ($p < .1$) association between gem-related disloyalty and gem fines in Experiment 2. I therefore suggest, given the plausibility of both these accounts given current and previous data, that both are likely to have played a role.

Surprisingly, neither children's judgements of transgression severity nor their 3PP severity were affected by deviations from group-descriptive norms, in contrast with previous studies that found that violators of descriptive norms are judged more severely than conformers by young children (Roberts, Gelman, et al., 2017; Roberts et al., 2018; Roberts, Ho, et al., 2017). However, in these studies children were presented with extremely simple scenarios in which characters were defined only by their adherence to descriptive norms. I thus propose that descriptive-norm related effects on children's moral behaviour, if present, might be weaker in more complex scenarios like the one adopted in Experiment 1, where descriptive information can be overridden by more salient injunctive information.

Another variable that did not affect children's 3PP severity was the presence of cues of observation. I suggest that, for a change in moral behaviour to be induced, cues of observation need to include the possibility that people will be held accountable for their own decisions. Such a possibility is absent in an anonymous experimental setting in which participants do not risk any obvious social cost for punitive behaviour or its absence (Raihani & Bshary, 2012). However, children's judgements of transgression severity proved to be more sensitive to cues of observation than punishment decisions. Indeed children expressed more negative judgements when being observed compared to when they were not being observed, mirroring the result Bourrat et al. (2011) found with adult participants. This evidence thus suggests that judgements of transgression severity and punishment severity decisions, although both involved in moral evaluation, are driven by distinct cognitive processes differentially affected by the very same cues of observation. It might also be that signalling higher moral outrage – rather than punishing transgressors more severely – is what confers more reputational benefits when in the presence of an observing audience.

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 (Dixson & Kenward, *in prep;* 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, consistent with the decrease in punishment severity between the ages of 5 and 11 I observed in Experiments 1-2. If it is indeed generally the case that rate of punishment increases with age, but punishment severity decreases, then it is likely that 3PP rates and severity follow distinct developmental trajectories with different cognitive underpinnings.

I now turn to my most unexpected and informative result – the vast majority of children showed no enjoyment of punishment, and even warning or pretending to punish was not enjoyed by most. Thus, contrary to my prediction, it is unlikely that retribution is a primary motivator of the observed 3PP. There are therefore two plausible explanations for the very high levels of punishment that were observed. Children may have been motivated by deterrence, or the demand characteristics of the experiment (taking the role of a referee) may have induced the children to think they were supposed to punish misbehaving players. In order 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 aforementioned lack of audience effects: perceived expectations to conform to the game norm might have already been close to ceiling in the no Audience conditions. Further, 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, López-Pérez, et al., 2019).

My research has a number of limitations that need to be acknowledged. First of all, a significant minority of children did not believe the moral scenarios they were refereeing had actually occurred. This did not affect the key variables I focused on but future work should aim at increasing realism of experimental settings. Believability issues, as well as the demand characteristics implicit in my 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 I manipulated the levels of observation (observed vs unobserved) 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. However, 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 I used. Finally, I 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.

Although 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. Further studies clarifying the potential roles of deterrence and conformity motivations for children's 3PP are now a priority. I will indeed return to this issue in Chapter 4 by providing a more in-depth examination of children's deterrent vs retributive motives.

Chapter 3: Experiment 3

CHILDREN'S INTENT-BASED MORALITY: A NON-WEIRD PERSPECTIVE

3.1. Introduction

3.1.1. General introduction

Experiment 3 went beyond Experiments 1-2 in that it is a modified version of the *MegaAttack* game that examined the extent to which children's moral judgements and decisions to enact 3PP are affected by information not only about players' behavioural outcomes but also about their underlying intentions. These analyses were conducted along the developmental trajectory and across different moral domains.

Additionally, I tested whether Experiment 2's finding with regard to lack of 3PP enjoyment could be generalised across different countries. Thus, including children from a less developed, more collectivistic Hispanic country (Colombia) was expected to greatly enhance the external validity of my previous work. More specifically, I reasoned that Experiment 2's evidence that children do not derive enjoyment from enacting 3PP with real (vs pretend) consequences on transgressors could be indicative of lack of a retributive motivation to punish. However, I could not rule out that children decided to punish under a retributive motivation expecting it to be satisfying, and when their expectations were not met they experienced low mood. In order to exclude this alternative explanation, in Experiment 3 it was crucial to investigate the temporal changes in children's 3PP enjoyment (in relation to the punishment time point). In Experiment 3 I also verified whether, provided that 3PP was presented as real, emphasising or not its consequences on transgressors would have an effect on children's 3PP enjoyment.

Moreover, in Experiment 2 it was found that children did not modify their punishment behaviour when subjected to an audience manipulation that cued observability. Therefore, in

Experiment 3 I adopted what I expected to be a more powerful audience manipulation – one that cued accountability in addition to observability. I then tested whether this could have an impact on children's 3PP severity. Finally, in Experiments 1-2 3PP severity was observed decreasing with children's increasing age, while controlling for judgements of transgression severity. Thus, a replication was attempted in Experiment 3 with the additional aim to check whether this developmental pattern was generalisable across different moral domains and cultures.

In summary, Experiment 3 was aimed at investigating the following issues in a small sample of children recruited from a collectivistic and developing Hispanic country: developmental patterns of children's morality across moral domains; how children's morality is modulated by an audience manipulation cuing accountability; how children's punishment-related affective states are affected by time and emphasis on punishment consequences on the transgressor; and the integration between outcome and intention information in children's morality along the developmental trajectory and across moral domains.

As the theory related to audience effects and punishment-related affective states has already been extensively covered in Chapter 2, in the next section I present only the theoretical background about the integration between intention and outcome information into children's moral judgement and behaviour.

3.1.2. Theoretical background

Moral evaluations rely on two crucial factors: the *outcomes* deriving from an agent's action and the agent's *intentions* behind such action. The integration of intentions and outcomes into both implicit and explicit moral evaluations is one of the milestones of children's development.

Infants, toddlers and very young children seem already impressively capable of integrating intentions and outcomes when implicitly evaluating moral agents, both in first-party evaluations (Behne, Carpenter, Call, & Tomasello, 2005; Dunfield & Kuhlmeier, 2010; Marsh, Stavropoulos, Nienhuis, & Legerstee, 2010; Vaish, Hepach, & Tomasello, 2018) and third-party evaluations (Chernyak & Sobel, 2016; Choi & Luo, 2015; Hamlin, 2013; Lee, Yun, Kim, & Song, 2015; Vaish et al., 2010; Woo, Steckler, Le, & Hamlin, 2017). Specifically, by using the unwilling vs unable paradigm, it has been demonstrated that infants show more signs of impatience when they interact with individuals that are unwilling rather than unable to give them a toy (Behne et al., 2005; Marsh et al., 2010). Infants prefer to help willing individuals over unwilling ones, even when such willingness does not translate into a positive outcome (Dunfield & Kuhlmeier, 2010). In a preferential reaching task, infants prefer intentional over accidental helpers, and accidental over intentional harmers (Woo et al., 2017). However, they do not show preferential reaching when the choice is between a successful helper and a puppet who intended to help but failed (Hamlin, 2013). By measuring spontaneous looking times, it has been established that around one year of age infants expect an agent to preferentially avoid intentional over accidental harmers (Choi & Luo, 2015), as well as hinderers over agents who intended to help, irrespective of whether they fail or succeed in their intent (Lee et al., 2015). Furthermore, there is evidence that 3-year-old children are more generous when reciprocating intentional rather than unintentional benefactors (Vaish et al., 2018), and more likely to correct adults' punishment decisions when they were imposed on puppets who committed accidental over intentional transgressions (Chernyak & Sobel, 2016). Additionally, children of the same age avoided helping adult actors who had harmful intentions, even when they did not result in harmful outcomes. Conversely, a selective avoidance is not detected towards actors who caused accidental harm (Vaish et al., 2010).

Whereas research employing spontaneous-response tasks has provided clear evidence that implicit moral evaluations are intent-based from a very early age, research adopting elicited-response tasks (i.e., series of questions after verbal story-telling, often accompanied by vignette presentation but lacking in behavioural cues of actors' intentions) with older children has produced a rather different picture (for a review, see Hilton & Kuhlmeier, 2019 and Margoni & Surian, 2016).

When children are asked to explicitly evaluate accidental and failed intentional transgressions, the presence of just one negative cue – either relating to outcomes or intentions – is sufficient to induce them to express a negative evaluation. However, differently from implicit moral evaluations, explicit ones do not appear to attribute more weight to intentions over outcomes until later in development. At a young age, children's explicit evaluations attribute equal weight to outcomes and intentions (Cushman, Sheketoff, Wharton, & Carey, 2013; Killen, Mulvey, Richardson, Jampol, & Woodward, 2011; Nobes, Panagiotaki, & Bartholomew, 2016), or more to outcomes over intentions (Costanzo, Coie, Grumet, & Farnill, 1973; Helwig, Hildebrandt, & Turiel, 1995; Helwig, Zelazo, & Wilson, 2001; Imamoğlu, 1975; Piaget, 1932/1965; Zelazo, Helwig, & Lau, 1996). It is only between 5 and 8 years of age – with the so-called "outcome-to-intent shift" – that explicit moral evaluations tend to become intent-based and that children's condemnation of accidental transgressions begins to decrease (Cushman et al., 2013). Regarding children's condemnation of failed intentional transgressions, there is both evidence that it remains steady (Cushman et al., 2013; Nobes, Panagiotaki, & Pawson, 2009) and also that it increases across development (Helwig et al., 1995).

What appears as a U-shaped developmental trajectory (i.e., moral evaluations being intent-based in infancy and toddlerhood, outcome-based in early childhood, again intent-based in middle childhood) has been accounted for by some scholars by making reference to the higher processing demands of elicited-response tasks compared to spontaneous-response tasks (Margoni & Surian, 2016). In elicited-response tasks, young children's limitations in executive functions (Zelazo et al., 1996) and explicit theory-of-mind skills (Killen et al., 2011) would prevent them from sophisticated consideration of intention information in their explicit moral evaluations. Thus, this view posits that the outcome-to-intent shift is determined only by cognitive changes outside the morality realm (*expression view*/*capacity model/parallel hypothesis*). This model also predicts that cognitive development, in terms of theory-of-mind and executive functioning, equally and simultaneously affects different types of moral evaluation. In other words, the outcome-to-intention shift would have the same onset for judgements of transgression severity and judgements of punishment acceptability. Additionally, across development reliance on intention information would increase for judgements of transgression severity as much as for judgements of punishment acceptability (Martin, Buon, & Cushman, 2019). By contrast, other scholars claim that the outcome-to-intent shift in explicit moral evaluations, although influenced by cognitive control resources, would be mostly due to a conceptual reorganisation inside the morality realm (*emergence view*/*theory model/constraint hypothesis*) (Margoni & Surian, 2016; Martin et al., 2019). Children's early moral concept – in which wrongness depends on negative outcomes – would be constrained by a later-developing one – in which wrongness depends on negative intentions. Different aspects of moral cognition would be affected differently by the development of this conceptual reorganisation. Specifically, the intent-based concept would nearly substitute the outcome-based concept in judgements of transgression severity, whereas the substitution would remain incomplete in judgements of punishment acceptability (Cushman et al., 2013). Consistent with this explanation, there is evidence that even during adulthood judgements of punishment acceptability tend to remain less reliant on intention information compared to judgements of transgression severity (Cushman, 2008, 2013; Cushman et al., 2013). Additionally, Cushman and colleagues argued that the substitution of the outcome-based concept with the intent-based concept would begin from judgments of transgression severity, and only later on would impact judgements of punishment acceptability. Although no clear theoretical explanations have been offered for this developmental lag, there is experimental evidence that the outcome-to-intent shift occurs in middle childhood for judgements of transgression severity, and in late childhood for judgements of punishment acceptability (Cushman et al., 2013). Moreover, a recent study has highlighted that the outcome-to-intent shift in actual third-party punishment (3PP) decisions becomes apparent even later than in judgments of punishment acceptability – in early adulthood (Gummerum & Chu, 2014).

3.1.3. Present study

Almost all literature about the role of intentions and outcomes in children's moral evaluations has been conducted in WEIRD (i.e., Western, Educated, Industrialised, Rich and Democratic) populations, particularly in English-speaking countries. Since studies in cognitive anthropology conducted on adult participants have provided evidence of remarkable cross-cultural variation in the weight of intentions (Barrett et al., 2016; Hamilton et al., 1983; McNamara, Willard, Norenzayan, & Henrich, 2019), it is expectable that WEIRD children are not necessarily representative of universal moral developmental patterns (Henrich et al., 2010). For this reason in my study I targeted non-WEIRD children, specifically urban Colombian children from low-middle socio-economic background. Colombia is a Western, Spanish-speaking, developing country with a very recent although fragile socio-political stability, following four decades of armed conflict. The ratification of the peace agreement between the government and revolutionary armed forces took place in 2016, but it has recently been challenged (Casey, 2016; Daniels, 2019).

To carry out my study on Colombian children I developed a variation of the *MegaAttack* game that had previously been employed to test British children (see Chapter 2). Regarding the content of the moral scenarios, I decided to represent only events in which outcomes and intentions had opposite valences, namely accidental transgressions (positive intention, negative outcome) and failed intentional transgressions (negative intention, positive outcome). These cases are the most informative to study how the relative weight of intentions and outcome changes with age. In addition to that, the fact that each video contained only one negative cue, either relating to outcomes or intentions, ruled out the potential inconvenience that children could merely anchor their moral evaluations to the first negative cue appearing in the scenarios (Nelson, 1980). Finally, the accidental and failed intentional transgressions shown in the scenarios represented, as in the experiments on British children, either a violation of fairness or disloyalty, respectively an individualising and a binding moral domain according to Moral Foundations Theory's definitions. On the basis of this theory, individualising moral domains include harm and fairness, and are defined as such because they are focused on the protection of individuals' rights. Loyalty, authority and purity instead constitute the so-called binding moral domains because they relate to the formation and maintenance of cohesive social groups (Graham et al., 2009). These distinctions appeared particularly relevant in light of prior work on US-based adults suggesting that the role of intentions varies across moral domains: it has been found that intentions matter more for moral judgements of harm, an individualising moral domain, and less for moral judgements of purity, a binding moral domain (Chakroff et al., 2015; Young & Saxe, 2011; Young & Tsoi, 2013).

3.1.4. Research hypotheses

The moral evaluations I took into consideration in Experiment 3 were of two types: judgements of transgression severity and 3PP severity decisions.

First of all, I expected Colombian children to express negative judgements and consequently to assign some punishment to both accidental and failed intentional transgressions. If the aforementioned *emergence view/theory model/constraint hypothesis* of moral development (Martin et al., 2019) were to be supported, also Colombian children would show a developmental lag between manifesting intent-based judgements of transgression severity and allocating intent-based punishments. The extent of this developmental lag might be similar to what has been detected in WEIRD populations, where it is by late preschool years that children begin to integrate intention information into their explicit judgments of transgression severity (Cushman et al., 2013), but it is not until early adulthood that this occurs for punishment severity decisions (Gummerum & Chu, 2014). Therefore, it was expected that Colombian children in the age range of choice (5-11 years of age) would show evidence of the outcome-to-intention shift in their judgements of transgression severity but not in their punishment severity decisions. This would translate into children judging failed intentional transgressions more severely than accidental transgressions, with this gap increasing with age (Helwig et al., 1995; Nobes et al., 2009). On the other hand, punishment severity assigned to failed intentional and accidental transgressions would be of roughly equal levels (Cushman et al., 2013), and this would remain consistent across the set age range (Gummerum & Chu, 2014). In other words, intentionality of the transgressions as well as the interaction between intentionality and children's age would be predictors of judgements of transgression severity but not of punishment severity.

Motivated by recent findings about the different weight of intentions across moral domains (Chakroff et al., 2015; Young & Saxe, 2011; Young & Tsoi, 2013), I investigated the relationship between the domain and intentionality of transgressions by taking into consideration both judgements of transgression severity and punishment severity decisions. I reasoned that, if the pattern of attributing more importance to the role of intentions in individualising over binding moral domains applies cross-culturally, Colombian children would assign more weight to intentions for judgements of unfairness severity compared to judgements of disloyalty severity. However, non-WEIRD populations tend to be more concerned about binding moral domains than WEIRD populations (Graham et al., 2013), therefore Colombian children might attach greater significance to intentions for judgements of disloyalty severity than for judgements of unfairness severity. I did not make any specific prediction regarding the effect of the interaction between domain and intentionality of transgressions on punishment severity.

I also explored whether some effects found in Experiments 1-2 with UK children (see Chapter 2) would conceptually replicate with Colombian children while using a different set of moral scenarios and/or manipulations. More specifically, in Experiment 3 I aimed to evaluate the effects of age and audience presence on Colombian children's judgement of transgression severity and punishment severity, controlling for intentionality and moral domain of the transgressions. I predicted that both judgement of transgression severity and punishment severity would decrease with increasing age as it had been observed both in Experiment 1 and 2 with UK children. I also explored whether the developmental patterns of judgement of transgression severity and punishment severity changed across moral domains, but I did not formulate any prediction. Moreover, by previously operationalising audience as the presence of cues of being observed, in Experiment 2 I found audience effects in judgements of transgression severity but not in punishment severity. I thus developed what I believed could be a stronger audience manipulation to verify whether it would exert any effects on punishment severity as well as judgements of transgression severity. My aim with the new audience manipulation was to activate in children not only concerns about being watched (like in Experiment 2) but also about being judged and held personally accountable for their actions (Raihani & Bshary, 2012). However, I did not commit to any prediction on that regard as this line of research has generated highly mixed results (Bradley, Lawrence, & Ferguson, 2018; Dear et al., 2019; Northover et al., 2017; Pfattheicher & Keller, 2015).

Additionally, prior research demonstrated that task framing could impact children's moral evaluations (Nobes et al., 2016; Zelazo et al., 1996). However, whether this could also be the case for children's punishment-related enjoyment was a completely unexplored question. Therefore, in Experiment 3 the framing manipulation was operationalised by varying the focus of the questions children were asked about their punishment-related enjoyment: children's decision to act as third-party punishers was framed in such a way to emphasise, or not, the outcomes for the transgressors which were caused by the punishment that the children imposed. I predicted that inducing children to focus on the consequences of their punishment decisions on the transgressors would decrease their enjoyment of punishment due to empathic concern towards the punished transgressors.

Furthermore, questions about children's punishment-related enjoyment were asked at three time points: 1) The first time point was before children's first punishment decision, namely children were required to predict how punishment would feel; 2) The second time point was immediately after children had punished for the first time; 3) The third and final time point was after children's last punishment decision, at the end of the experiment. My previous results from a UK sample suggested children generally did not derive enjoyment from punishing when enquired about their affective states at the end of the experiment (see Experiment 2, Chapter 2), therefore in the current study I expected to find a similar result for the same time point. I additionally predicted that Colombian children's enjoyment levels would vary according to the timing of the enjoyment question: the direction of this effect would shed further light of children's punishment motivations. If Colombian children anticipate punishment to feel worse or as bad as how it actually feels at the second and third time points, this would be indicative of lack of a retributive motivation to punish. Conversely, if Colombian children expect punishment to be satisfying, this would constitute indication of a retributive motivation. Depending on whether their retributive expectations are subsequently met or not, children's mood would remain positive or, as it was observed by Carlsmith, Wilson and Gilbert (2008) in US adults, would lower at the second and third time points.

3.2. Method

3.2.1. Materials

The *MegaAttack* game was run on a laptop computer which was taken to the test location. In the test trials, participants saw recordings of games that they were told were being played and commented on live by internet players. As in Experiment 2, players in Experiment 3 were supposed to equally distribute bombs between each other, collect normal sized-gems while defending themselves from enemies' attacks, and participate in a cooperative task for the collection of a mega-gem. This computerised paradigm gave me the opportunity to present moral scenarios in such a way that children could infer characters' intentions by observing their behaviour and listening to their dialogues as opposed to the experimenters representing their mental states. Each video presenting the moral scenarios via game bouts was kept short (~1 minute each) with the aim of not excessively taxing children's working memory. Questions being asked to the children did not require articulated verbal responses. All these precautions were made to minimise the cognitive demands of my elicited-response task (Armsby, 1971; Farnill, 1974; Hilton & Kuhlmeier, 2018; Yuill, 1984; Yuill & Perner, 1988).

3.2.2. Sample

Participants were 44 primary school-aged Colombian children (*mean age*: 7.90 ± 1.34 years; *age range*: from 5.83 years to 10.84 years; 12 females and 32 males). Children were tested in one primary school of the capital, from July 2018 to March 2019. The stopping rule was to collect as much data as possible by the end date of my gatekeeper's working contract. The study was approved by Los Andes University Ethical Review Committee and later received Chair's approval by Oxford Brookes University.

All caregivers (31 biological mothers; 7 biological fathers; 3 grandmothers; 1 adoptive father; 1 aunt; 1 stepmother) partially or fully completed a socio-demographic questionnaire, indicating that Experiment 3's sample was all of Colombian nationality, of low socio-economic status with a low-middle education level (11% of the respondents had a primary school qualification; 57% a secondary school qualification; 23% a post-secondary school technical qualification; 9% a Bachelor's degree).

3.2.3. Design

I adopted a mixed design in which the factors were: *Domain of moral transgression* (2 within-subject levels: unfairness; disloyalty); *Intentionality of moral transgression* (2 within-subject levels: failed intentional attempt; accident); *Enjoyment question timing* (3 within-subject levels: before; during; after); *Enjoyment question focus* (2 between-subject levels: outcome of punishment; no outcome of punishment); *Audience* (2 between-subject levels: present; absent).

Order with respect to failed intentional/accidental transgression was ABBA or BAAB, and with respect to disloyalty/unfairness transgression was ABAB or BABA, counterbalanced (see Appendix B – Table β1). Each of the four resulting test trials featured a different pair of player avatars (different animals inside space-ships).

The dependent variables measured were: *Judgement of transgression severity* (using the same 11-point scale as in Experiment 2); *Severity of punishment* (6 ordinal levels ranging from 1, “no punishment”, to 6, “1 day-ban”, as in Experiments 1-2 but with just one punishment type – social); *Punishment-related enjoyment* (as in Experiment 2 it had 11 ordinal levels from -5, “very bad”, to +5, “very good”);

3.2.4. Procedure

The procedure of Experiment 3 closely resembled that of Experiment 2, thus this section mostly highlights the differences. **Playing familiarisation** was identical to Experiment 2. Between the playing and refereeing familiarisations there was a **refereeing introduction**, differently characterised depending on whether children had been assigned to Audience or No Audience condition (see script in Appendix B – section β1.4). **Refereeing familiarisation** was identical to Experiment 2 for children in the No Audience condition; for children in the Audience condition there were some additional features (see further details about the audience manipulation below in this section).

The refereeing familiarisation was followed by **four test trials** in which the child saw a combination of moral transgressions varying in terms of domain and intentionality (see Appendix B – Table β1). During the test trials children did not see any live-stream commentator, differently from Experiment 2. In place of that, children could hear purportedly live dialogues (actually pre-recorded) between the internet players; gender of these voice-overs was matched with that of the child being tested. Regarding the moral transgressions being shown in the videos, they could be either accidental transgressions or failed intentional transgressions, related either to the fairness or loyalty domain. Accidental transgressions were characterised by players having positive intentions (to be fair or loyal), followed by negative outcomes (unfair distribution of resources or failure to help in the mega-gem cooperative task, like in Experiment 2). Conversely, failed intentional transgressions were characterised by negative intentions (to be unfair or disloyal) followed by positive outcomes (fair distribution of resources or participation in the mega-gem cooperative task). Specifically, in the case of accidental unfairness one player intended to split the bombs (i.e., defensive resources) equally with the team-member (5 bombs each, out of 10) but, by mistake, ended up with more bombs (7/10) than the equal share. In the case of failed attempt at unfairness, one player intended to take for themselves more bombs (7/10) than the equal share, but inadvertently ended up allocating equal numbers of bombs (5/10) to themselves and the team-member. In instances of accidental disloyalty one player intended to cooperate with the team-mate in the mega-gem collection but, due to a mistake, did not succeed in freeing the team-mate from the mega-gem (who thus remained exposed to enemies' attacks). In failed attempts at disloyalty one player intended to leave the team-mate trapped in the mega-gem, but inadvertently set them free from the trap.

After having seen each of the five internet scenarios (1 refereeing familiarisation + 4 test trials), the child had to answer for each of the two players in turn: "*Did this player behave badly?*". If a misbehaviour was identified, the child had to provide a judgement of transgression severity (same Likert scale as in Experiment 2) and decide the consequent punishment severity (same Likert scale as in Experiment 2). Differently from Experiment 2, in Experiment 3 children did not have to choose the type of punishment: they had at their disposal only social punishment in the form of a ban from the game (with the same options for Experiment 2's social punishment, inclusive of "no punishment" option).

Whereas in Experiment 2 I adopted a between-subjects manipulation to explore children's punishment-related enjoyment (three punishment-opportunity conditions: real punishment; warning about possible future punishment; pretend punishment), in Experiment 3 the measure of punishment-related enjoyment was conducted within-subjects. Specifically, here children were asked to provide their enjoyment rating at three time points (same Likert scale as in Experiment 2): 1) before their first punishment decision by forecasting how punishment would feel (time point: before); 2) once they had punished for the first time (time point: during); 3) after the last punishment decision (time point: after). The focus of such questions about punishment-related enjoyment was varied between subjects: the framing of the questions highlighted, or not, the outcomes for the transgressors following children's administration of punishment (outcome vs no outcome focus). For example, at the first time point children in the outcome focus condition were asked: "*So, you might ban some players from the game so they can't play for quite a while. How do you think it will feel to do that?*". Instead, children in the no outcome focus condition were simply asked: "*So, you might punish some players. How do you think it will feel to do that?*" (for the details of the framing manipulation for each time point see Appendix B – sections β1.4, β1.7, β1.8).

Regarding the audience manipulation, the audience cues I adopted were completely different from Experiment 2. Indeed, the attention of the experimenter was constant across audience conditions; internet players' voice-overs were always audible; the frame outside the game arena was always empty (no player-avatars observing); children were always identifiable via the nickname of their refereeing avatar, independently of audience condition. My aim was to make children feel they were not merely being observed, but that they were being judged and held accountable for their punishment decisions. To do that, in the refereeing introduction children being assigned to the Audience condition were told their refereeing performance would be rated by a referee chosen from the highest scoring referees in the *MegaAttack* game championships, who would act as their mentor. The leader-board with the names of the best referees was shown to the children in the Audience condition two times: before the refereeing familiarisation and half away through the test trials (as a reminder). Meanwhile, children in the No Audience condition were told no-one would evaluate their refereeing decisions; they did not hear any mention of the *MegaAttack* game championships and consequently were never shown the leader-board. Afterward, once children in the Audience condition were allocated a specific referee-mentor at the beginning of the refereeing familiarisation, the experimenter drew the children's attention towards the presence of the mentor's avatar. At the end of each internet video (both in the refereeing familiarisation and in the test trials), the mentor's avatar started moving and the experimenter reminded the child this was the signal to indicate the mentor was going to judge their refereeing decisions. Finally, after the end of the game (i.e., after all video trials) children in the Audience condition were shown the score their mentor purportedly gave to their refereeing performance. For ethical reasons, all the children received the same score, namely 10 out of 10.

At the **end of the experiment**, each child was asked two manipulation check questions. The manipulation check questions were to evaluate the effectiveness of the audience manipulation and the believability of the experimental setting. Specifically, children were questioned about whether they felt their refereeing decision were being judged ("*Did you feel like your decisions as a referee were judged by others?*"), and whether they thought they had actually refereed real internet players during the trials ("*Do you think you really watched games with internet players now?*").

3.2.5. Analysis Strategy and Statistics

To test my research hypotheses, I adopted linear models implemented using the lme4 package (Version 1.1-21) in the R programming environment (Version 3.5.1, R Core Team, 2018). Depending on the type of dependent variables, I used different types of functions: *lmer* to conduct linear mixed-effects analyses on continuous variables (judgement of transgression severity; punishment severity; punishment-related enjoyment), and *glmer* to conduct generalised linear mixed-effects analyses on binary variables (punishment-related enjoyment). To note, punishment-related enjoyment was analysed in two ways: as a continuous variable (with values ranging from -5 = “very bad” to +5 = “very good”) and then, for comparability with my previous work (Experiment 2, Chapter 2), also as a binary variable (values greater than 0 were recoded as “enjoyment”, while values less than or equal to 0 were recoded as “no enjoyment). All explanatory models included a range of independent variables; all but one were used as fixed factors (details in the tables presented in section 3.3.2). Only children’s ID was used as a random factor since there were multiple data points per individual. The significance of fixed and random effects were obtained via ANOVA tests of the full model with the effect in question against the model without the effect in question.

The statistics I utilised to calculate effect sizes is based on Nakagawa, Johnson, and Schielzeth’s (2017) paper. I obtained two main values for each model: $R^2_{GLMM(m)}$, which measures fit of the fixed components of the model, and $R^2_{GLMM(c)}$, which measures fit for fixed and random components together. Effect sizes for individual fixed factors were stated as ΔR^2, defined as the reduction in $R^2_{GLMM(m)}$ when that factor was removed from the model, but all other factors remained.

3.3. Results

3.3.1. Preliminary analyses

3.3.1.1. Audience manipulation

When children were assigned to the No Audience condition, they felt their decisions as referees were not being judged in 18 out of 21 cases, thus their comprehension rate was 86%, 95% CI [64%, 97%]. When instead children were assigned to the Audience condition, they felt they were being judged by others only in 12 out of 22 cases, thus their comprehension rate was 55%, 95% CI [32%, 76%]. A Fisher’s exact test revealed that the comprehension rate was significantly higher in the No Audience condition than in the Audience condition, *p* = .010.

Because of the low comprehension rate in the Audience condition, in the models I developed for the Main Analyses (see Tables 2-3 and 6-7) I entered Audience perception rather than actual Audience condition as a modulating factor.

3.3.1.2. Believability of the game

A large majority of children (40 out of 44) believed they had refereed real games, thus the level of believability was 91%, 95% CI [77%, 97%]. There was no effect of believability on the key variables (see Tables 2-3 and 6-7).

3.3.1.3. Recognition and punishment rate

Misidentifications of non-transgressors as transgressors were made by 3 children, for a total of 4 trials, all in the refereeing familiarisation. Refereeing familiarisation trials were anyway not included in the analyses.

Regarding the test trials, children correctly recognised the transgressors in 138 out of the total 176 times a moral transgression was shown, thus the recognition rate was 78%, 95% CI [71%, 84%]. The recognition rate was affected by both the domain (χ^2 (3) = 12.15, p = .007, $\Delta R^2 = .142$) and intentionality of transgressions (χ^2 (3) = 9.40, p = .024, $\Delta R^2 = .125$). Acts of disloyalty were correctly recognised as transgressions (M = 88%, SD = 33%) more frequently than acts of unfairness (M = 69%, SD = 46%). Failed intentional attempts were recognised as transgressions (M = 84%, SD = 37%) more frequently than accidents (M = 73%, SD = 45%).

In 129 out of the 138 times a moral transgression was correctly recognised children decided to punish the transgressor, thus the punishment rate was 93%, 95% CI [88%, 97%].

3.3.2. Main analyses

3.3.2.1. Judgement of transgression severity in Experiment 3

Linear mixed-effects analyses conducted using the model I developed to explain judgement of transgression severity (Table 2) revealed a significant interaction between moral domain and intentionality of the transgressions (Figure 5): whereas judgements of unfairness were comparable in terms of severity between accidents (M = -3.41, SD = 1.82) and failed intentional attempts (M = -3.45, SD = 1.62), instances of failed intentional disloyalty were judged more severely (M = -4.19, SD = 1.37) than those of accidental disloyalty (M = -3.17, SD = 1.84). Furthermore, children's age as well as interactions between age and domain or between age and intentionality of the transgressions were not predictors of judgements of transgression severity. Finally, the perceived presence of an audience did not have any effect on judgement of transgression severity (audience present: M = -3.48, SD = 1.69; audience absent: M = -3.72, SD = 1.64).

Table 2. **Modulating factors of judgement of transgression severity in Experiment 3.**

Factor	χ^2	p	ΔR^2
Age	0.72	.869	.004
Gender	0.19	.659	.003
Audience perception	0.09	.759	.001
Believability	0.07	.789	.001
Moral domain	5.81	.121	.024
Intentionality	10.85	.013*	.047
Moral domain x Intentionality	4.18	.041*	.016
Age x Moral domain	0.38	.538	.002
Age x Intentionality	0.17	.677	.001

* $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$.

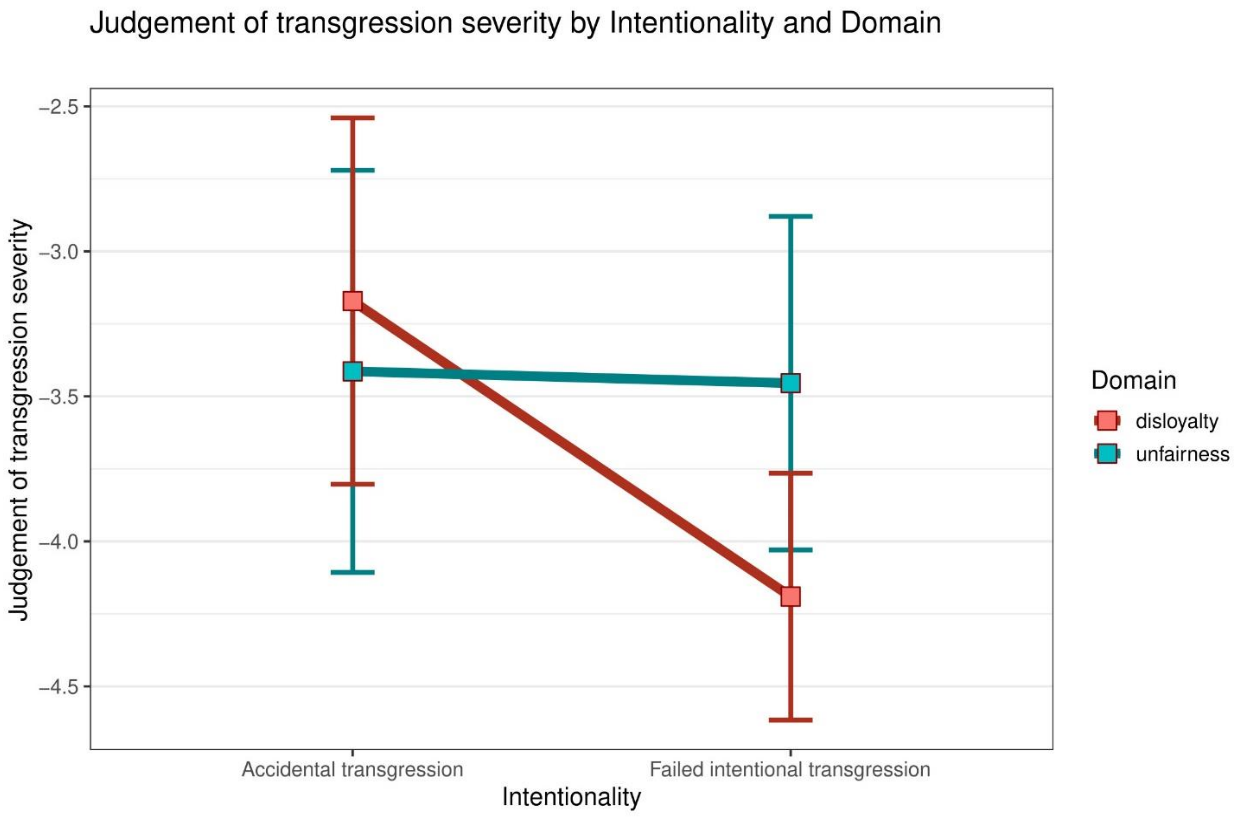

***Figure 5.* Judgement of transgression severity by Intentionality (accidental transgression vs failed intentional transgression) and Moral domain (disloyalty vs unfairness) in Experiment 3.**

3.3.2.2. Punishment severity in Experiment 3

Linear mixed-effects analyses conducted on the model I developed to explain punishment severity (Table 3) did not reveal any significant interaction between moral domain and intentionality of the transgressions (Figure 6). Punishment severity was comparable between accidental (M = 3.67, SD = 1.52) and failed intentional transgressions (M = 4.05, SD = 1.54). However, the moral domain of the transgression was a significant predictor: children punished instances of disloyalty more severely (M = 4.03, SD = 1.56) than of unfairness (3.69, SD = 1.51). Moreover, children's increased age significantly predicted decreased punishment severity, B = -0.51. Importantly, however, there was a significant interaction between the moral domain of the transgression and children's age in predicting children's punishment severity. This means that punishment severity decreased with increasing age only in cases of unfairness (irrespective of transgressor's intentionality), whereas it remained stable across ages in cases of disloyalty (again irrespective of intentionality) (see Figure 7). Finally, the perceived presence of an audience did not exert any effect on punishment severity (audience present: M = 3.63, SD = 1.57; audience absent: M = 4.01, SD = 1.53). Judgement of transgression severity was a significant predictor of punishment severity: the more severely children judged the transgressions, the harsher they were in punishing the transgressors, B = - 0.20.

***Table 3*. Modulating factors of punishment severity in Experiment 3.**

Factor	χ^2	***p***	ΔR^2
Judgement of transgression severity	21.17	< .001 ***	.140
Age	13.43	.004 **	.083
Gender	0.32	.572	.002
Audience perception	2.37	.123	.018
Believability	2.66	.103	.020
Moral domain	10.03	.018 *	.057
Intentionality	3.21	.360	.017
Moral domain x Intentionality	0.42	.516	.003
Age x Moral domain	8.31	.004 **	.047
Age x Intentionality	1.31	.253	.007

* $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$.

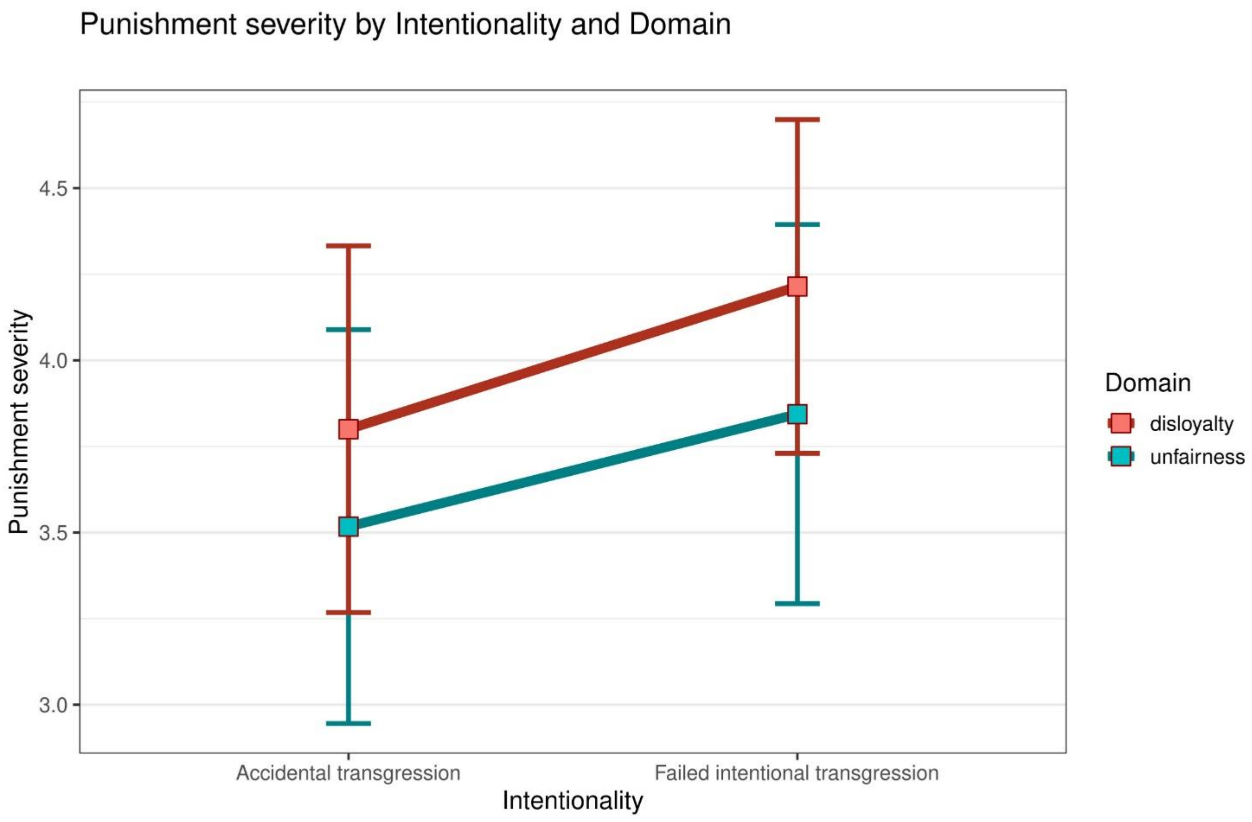

Figure 6. **Punishment severity by Intentionality (accidental transgression vs failed intentional transgression) and Moral domain (disloyalty vs unfairness) in Experiment 3.**

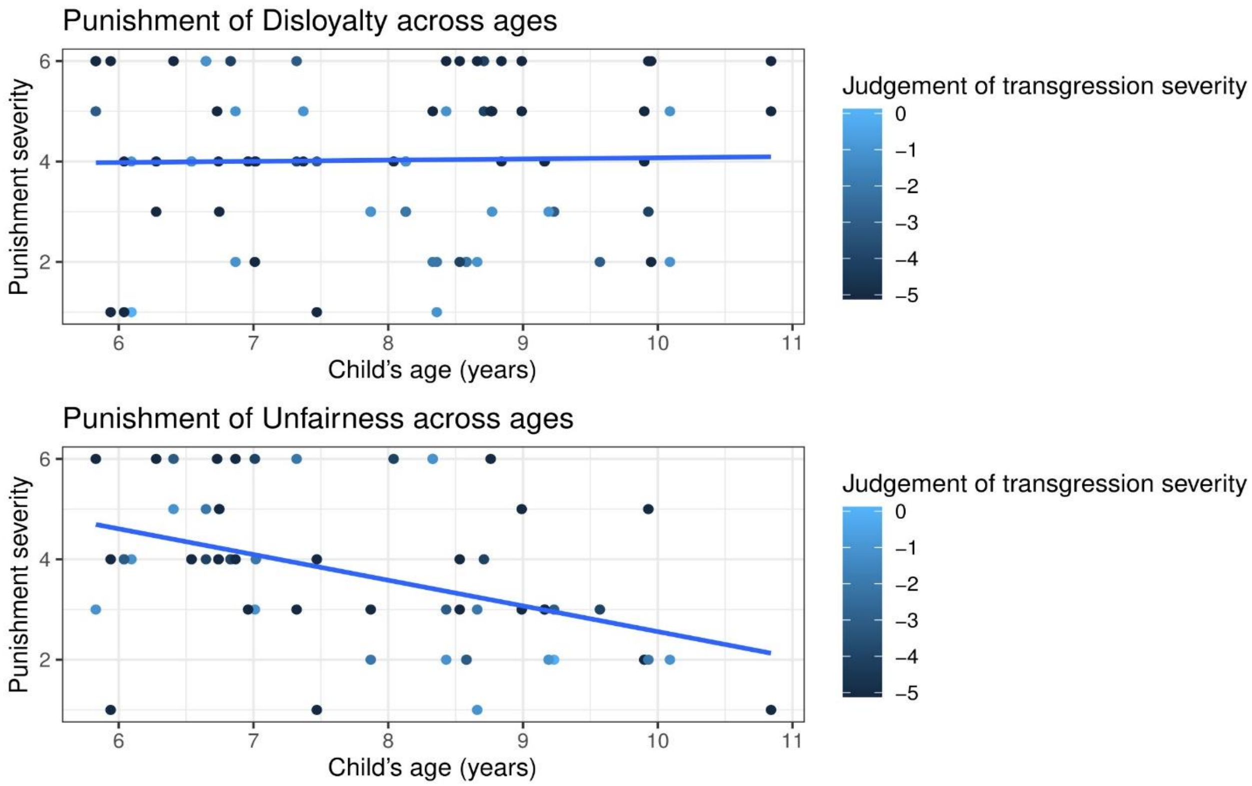

***Figure 7.* Developmental pattern of punishment severity by moral domains (disloyalty vs unfairness) in Experiment 3, with reference to judgement of transgression severity.** Judgement scale ranging from -1 = "just a little bad" to -5 = "very very bad".

3.3.2.3. Comparisons with punishment severity in previous experiments

Interestingly, it was found that Colombian children's punishment severity was predicted not only by main effects of domain and age but also by an interaction between domain and age. This prompted my interest in re-analysing the dataset of my previous two experiments (similarly aged British children tested with a variation of the *MegaAttack* game) to see whether these patterns replicated or not across UK and Colombia and these results are presented below.

3.3.2.3.1. Reanalysis of Experiment 1 with respect to predictors of punishment severity

In Experiment 1, where the punishment could mitigate unfairness transgressions (i.e., gem fine for gem unequal distribution), linear-mixed effects analyses revealed that British children's punishment severity was significantly predicted by age, moral domain and the interaction between age and domain, while controlling for judgements of transgression severity (see Table 4). Specifically, children's age significantly predicted decreased punishment severity, B = -0.21. Acts of unfairness were punished more severely (M = 4.44, SD = 1.09) than acts of disloyalty (M = 4.23, SD = 1.23). Interestingly, British children's punishment severity decreased with increasing age only in cases of disloyalty, whereas it remained stable across ages in cases of unfairness (see Figure 8), thus showing the opposite interaction to what was detected with Colombian children in Experiment 3.

***Table 4*. Modulating factors of punishment severity in Experiment 1.**

Factor	χ^2	p	ΔR^2
Judgement of transgression severity	20.40	< .001 ***	.069
Age	14.06	.003 **	.075
Gender	0.37	.542	.001
Believability	1.12	.572	.010
Moral domain	10.38	.016 *	.017
Descriptivity	0.83	.841	.001
Moral domain x Descriptivity	0.38	.538	.001
Age x Moral domain	6.22	.013 *	.010
Age x Descriptivity	0.35	.555	.000

* $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$.

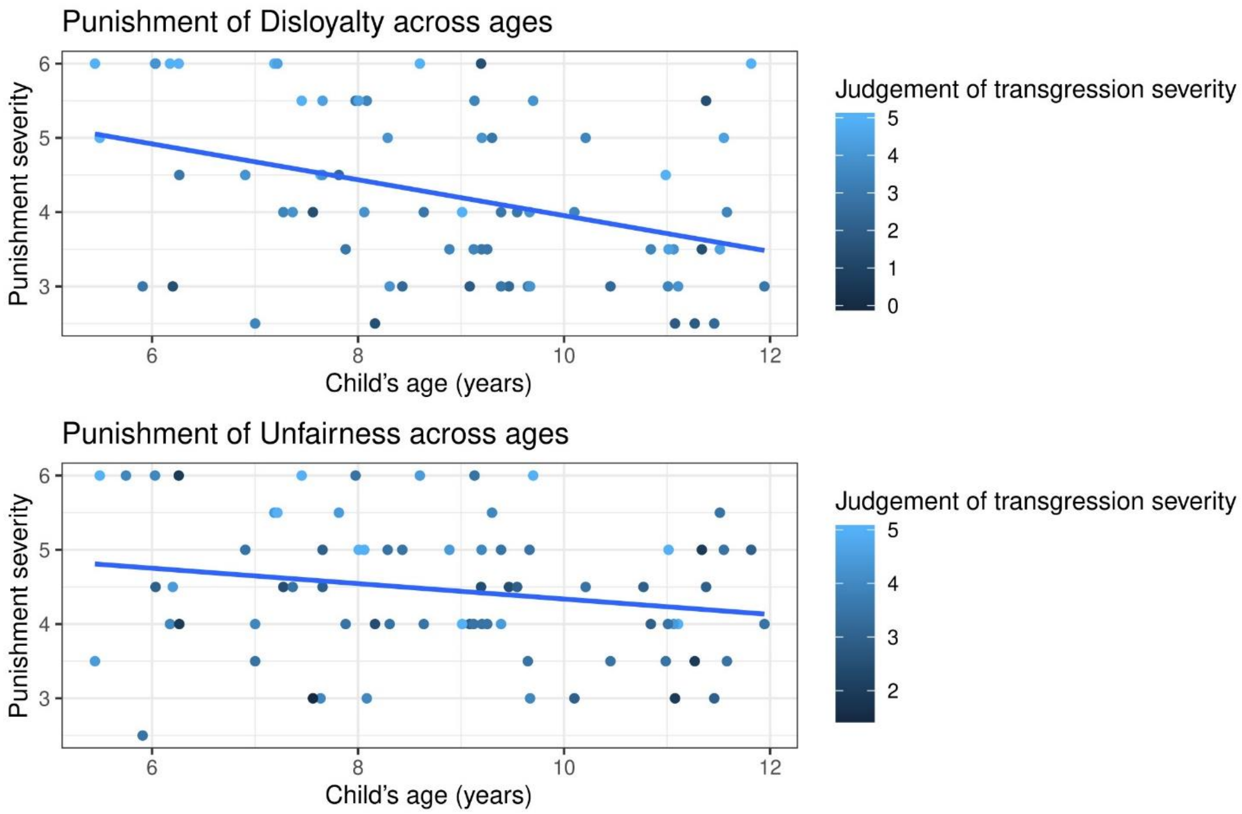

Figure 8. **Developmental pattern of punishment severity by moral domains (disloyalty vs unfairness) in Experiment 1, with reference to judgement of transgression severity.** Judgement scale ranging from 0 = "not bad" to 5 = "super bad".

3.3.2.3.2. Reanalysis of Experiment 2 with respect to predictors of punishment severity

Differently from Experiment 1, in Experiment 2 the punishment types children could allocate were not designed to reduce the inequity caused by unfairness transgressions. In this case, linear-mixed effects analyses revealed that British children's punishment severity was significantly predicted by age, but not by domain or by the interaction between age and domain, again while controlling for judgements of transgression severity (see Table 5). Therefore, it could be concluded that punishment severity decreased with increasing age both in cases of unfairness and disloyalty, B = -0.24, in contrast with both Experiment 1 and Experiment 3. Punishment severity for acts of disloyalty (M = 4.47, SD = 1.34) was comparable to that for acts of unfairness (M = 4.33, SD = 1.44), see Figure 9.

***Table 5*. Modulating factors of punishment severity in Experiment 2.**

Factor	χ^2	p	ΔR^2
Judgement of transgression severity	15.12	< .001 ***	.046
Age	20.92	< .001 ***	.132
Gender	3.68	.055	.022
Believability	4.53	.104	.025
Moral domain	6.25	.100	.010
Audience	3.25	.354	.003
Moral domain x Audience	3.02	.082	.003
Age x Moral domain	3.27	.071	.007
Age x Audience	< .001	.988	.000

* $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$.

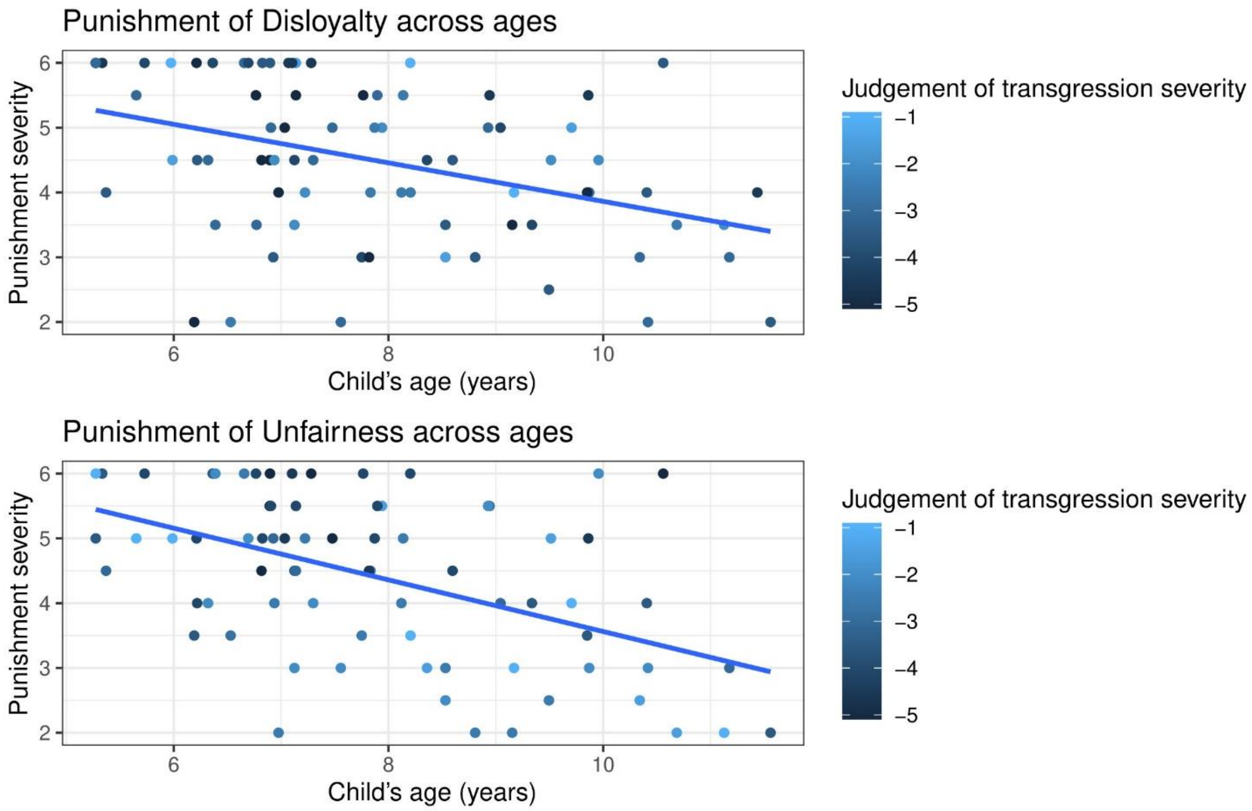

***Figure 9.* Developmental pattern of punishment severity by moral domains (disloyalty vs unfairness) in Experiment 2, with reference to judgement of transgression severity.** Judgement scale ranging from -1 = "just a little bad" to -5 = "very very bad".

3.3.2.4. Punishment-related enjoyment in Experiment 3

On average children did not enjoy punishing: across conditions M = -0.89, SD = 2.82, which was significantly more negative than 0, t(43) = -2.08, *p* = .043, d = -0.31, 95% CI for d [-0.62, -0.01].

I first ran linear mixed-effects analyses on the model developed to explain punishment enjoyment as a continuous variable (Table 6). It was found that children's punishment enjoyment did not depend on whether the punishment enjoyment question being asked was focused on punishment outcomes (M = -0.33, SD = 3.26) or not (M = -1.31, SD = 2.42). Likewise, the analyses also showed that the timing of the punishment enjoyment question (before; during; after punishment allocation) approached but did not reach significance as a predictor of punishment-related enjoyment (before-enjoyment: M = -1.64, SD = 3.12; during-enjoyment: M = -0.53, SD = 3.52; after-enjoyment: M = -0.57, SD = 3.93) (Figure 10).

I subsequently ran generalised linear mixed-effects analyses on the model developed to explain punishment enjoyment as a binary variable (Table 7). It could be confirmed that the focus of the enjoyment question did not exert any effect on whether children enjoyed punishment (enjoyment score > 0) or not (enjoyment score ≤ 0). Conversely, it was found that enjoyment question timing did exert a significant effect on whether children enjoyed punishment or not. Specifically, the percentage of participants that reported no enjoyment was 84% when children forecasted how punishment would feel (time point: before); 67% once children had punished for the first time (time point: during); 61% after the last punishment decision (time point: after). Post-hoc paired comparisons (Mc Nemar's tests) revealed that the differences between before-enjoyment and during-enjoyment as well as between before-enjoyment and after-enjoyment were significant (respectively, *p* = .035 and *p*

= .012). Instead, the difference between during-enjoyment and after-enjoyment was not significant, p = .366 (Figure 10).

***Table 6.* Modulating factors of punishment-related enjoyment as a continuous variable in Experiment 3.**

Factor	**χ^2**	***p***	**ΔR^2**
Judgement of transgression severity (average)	0.04	.842	.001
Age	2.22	.136	.030
Gender	0.76	.384	.010
Audience perception	0.02	.876	.000
Question focus (outcome vs no outcome)	1.40	.238	.018
Time (before; during; after)	5.58	.062	.024
Believability	0.35	.556	.004

* $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$.

***Table 7.* Modulating factors of punishment-related enjoyment as a binary variable (enjoyment vs no enjoyment) in Experiment 3.**

Factor	**χ^2**	***p***	**ΔR^2**
Judgement of transgression severity (average)	0.76	.384	.004
Age	7.24	.007 **	.091
Gender	1.83	.176	.024
Audience perception	0.04	.840	.001
Question focus (outcome vs no outcome)	2.28	.131	.029
Time (before; during; after)	9.58	.008 **	.086
Believability	0.05	.830	.001

* $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$.

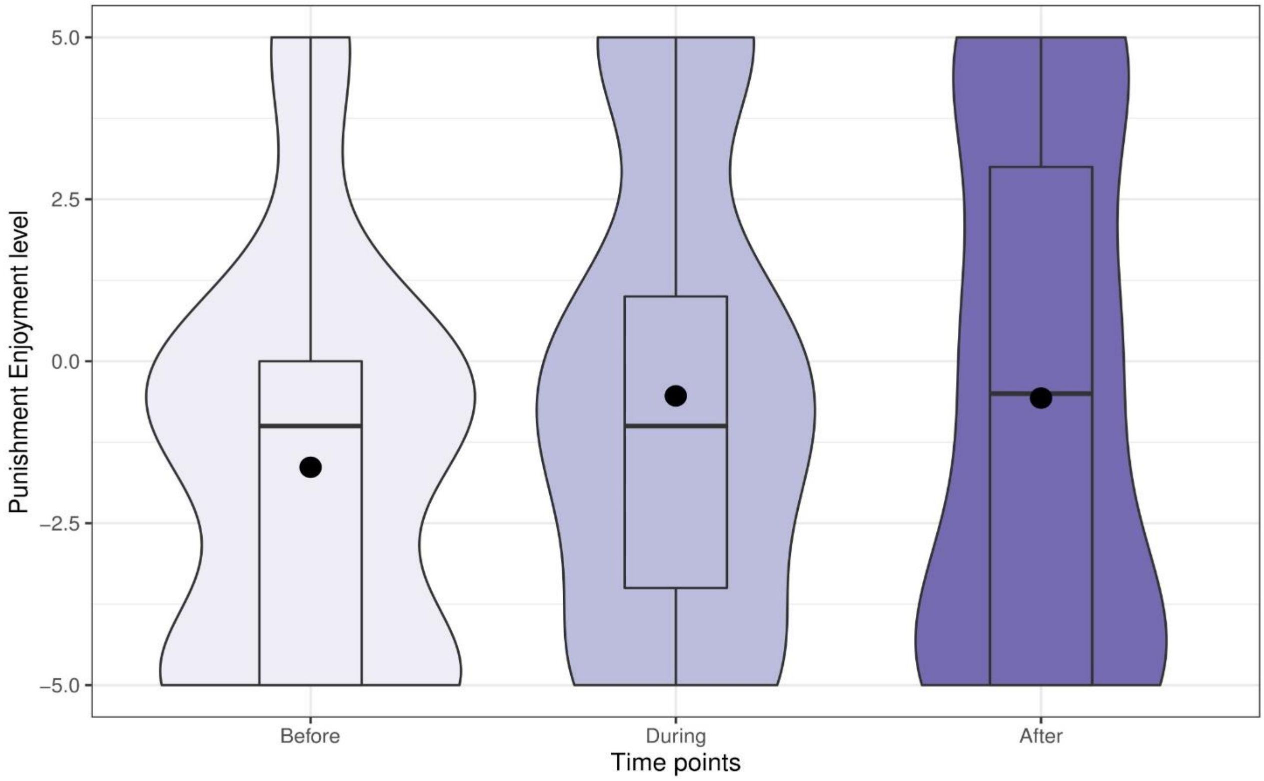

***Figure 10.* Punishment-related enjoyment by time points of Experiment 3: before; during; after punishment allocation.** Violin plots wrapping boxplots; boxplots showing median and interquartile range, outliers, and a large dot for mean value.

3.4. Discussion

My study has expanded the knowledge about the role of outcomes and intentions in children's judgements of transgression severity, punishment severity and punishment-related enjoyment by testing a non-WEIRD sample, namely 5- to 11-year-old urban Colombian children from a low-middle socio-economic background.

I replicated Cushman et al.'s (2013) and Gummerum and Chu's (2014) findings by demonstrating that in this age range intentionality of the transgressions exerted a significant effect on judgements of transgression severity but not on actual decisions about punishment severity. On average, failed intentional transgressions were indeed judged more severely than accidental transgressions, whereas failed intentional transgressions were punished as severely as accidental transgressions. This intentionality mismatch between the two types of moral evaluations indicates that even in Colombian children – as in WEIRD children – the onset of the outcome-to-intention shift in judgements of transgression severity occurs prior to that in punishment severity decisions. Specifically, signs of the former shift were detected already in 5-year-olds, while no signs yet of the latter shift were present in 11-year-olds. The fact that the occurrence of the outcome-to-intent shift in Colombian children was not simultaneous across judgements of transgression severity and punishment severity decisions is thus in contrast with the *expression view/capacity model/parallel hypothesis* (Margoni & Surian, 2016) but in accordance with the *emergence view/theory model/constraint hypothesis* (Martin et al., 2019). This suggests that developmental changes in children's explicit moral evaluations are primarily the consequence of cognitive changes inside (i.e., conceptual reorganisation) rather than outside (i.e., executive functions and explicit theory-of-mind skills) the realm of morality.

Interestingly, the outcome-to-intention shift in judgements of transgression severity was partially moral domain-dependent. Judgements of unfairness were of equal severity between accidental and failed intentional transgressions, while judgements of disloyalty were much harsher for failed intentional than accidental transgressions. This means that Colombian children considered intentions more important than outcomes for judgements related to a binding moral domain (disloyalty) than for judgements related to an individualising moral domain (unfairness), thus reacting in the opposite way compared to WEIRD adults (Chakroff et al., 2015; Young & Saxe, 2011; Young & Tsoi, 2013). Additionally, Colombian children were shown to punish disloyalty more severely than unfairness transgressions; interestingly, punishment severity for unfairness (but not disloyalty) transgressions decreased with children's increasing age. Importantly, these results were obtained while controlling for judgements of transgression severity (which was not influenced by age, differently from Experiments 1-2). Conversely, 5- to 11-year-old UK children being tested on a similar paradigm reacted to the view of transgressions in quite a different way. When UK children could use punishment not only to make the transgressor pay for their action but also to equalise the unbalance between victim and transgressor, they proved to be especially concerned about fairness over loyalty. This is testified by the fact that punishment severity of unfairness was higher than that of disloyalty and remained stable across ages while punishment severity of disloyalty decreased (Experiment 1, Chapter 2). When, instead, UK children could not use punishment for equalisation purposes (thus resembling the experimental setting Colombian children were confronted with), their punishment severity was comparable across moral domains and decreased with an age-dependent pattern for instances of disloyalty and unfairness alike (Experiment 2, Chapter 2). My findings are thus in line with Moral Foundations Theory's (Graham et al., 2013) argument that non-WEIRD cultures are particularly concerned about binding over individualising moral domains when evaluating moral behaviours, and that such selective concerns become more pronounced with development because of culture-directed learning processes. Whereas Moral Foundations Theory offers an explanation for cultural differences in the relative weight of moral domains *per se*, cultural group selection (Richerson & Boyd, 2005) can explain why intentionality is more important within the moral domains privileged by a specific culture. Since negative intentions are a stronger predictor of recidivism than accidents, it makes evolutionary sense that people are watchful about clues indicating someone's intention to disregard the moral norms their own group cares particularly about (e.g., loyalty norms in collectivistic societies). However, I anticipate this general pattern to have exceptions in case of irreparable crimes, whose severity is already at ceiling even in the absence of recidivism (e.g., desertion, homicide). In those circumstances I would expect a negative relation between the importance of transgression domain and the importance of intentionality within it. In other words, when a heinous crime is committed, people are likely to stop caring about whether such act was deliberate or not.

There were some unexpected null results in Experiment 3 which merit discussion. It was found that, as children's age increased, neither did judgements of accidental transgressions become more lenient nor did judgements of failed intentional transgressions become harsher. These results contrast with findings of previous studies (Cushman et al., 2013; Helwig et al., 1995; Nobes et al., 2009) where, for judgements of transgression severity, the weight of intentions relative to outcomes increased with increasing age starting from approximately 5 years of age. I speculate that, whereas the onset of the outcome-to-intention shift occurs with a similar timing in WEIRD and non-WEIRD children, the increase in the gap between judgements of accidental and failed intentional transgressions might have a more culture-dependent progression: faster and continuous in WEIRD children; slower and in stages in non-WEIRD children. In order to explore this possibility, future research targeting non-WEIRD populations should test elementary school-aged children with a comparison group of adolescents. This would also allow the scientific community to understand whether the difference between judgements of disloyalty vs unfairness severity (if replicated) remains steady with development. Alternatively, this null result might be simply explained by the low processing demands of the task I used, which could have allowed even 5-year-old children to already show a mature moral competence in weighing intentions into their judgements of transgression severity.

Another variable that did not affect either children's judgements of transgression severity or 3PP severity was audience, precisely the perceived presence of cues of accountability. This is at odds with previous findings that showed judgements of transgression severity being harsher in the presence of "watching eyes" (adult sample: Bourrat et al., 2011) or other cues of observability (children sample: Experiment 2, Chapter

2). An explanation for the absence of audience effects to consider is that people tend to base their moral judgements and punishment decisions on the heuristic that reputation is generally at stake, even when an audience judging their actions is factually absent (Jordan & Rand, 2019; Tennie, 2012). This implies that even when they did not consciously perceive any audience holding them accountable, children in Experiment 3 might have been motivated to increase the harshness of their judgements and punishment decisions with the (non-conscious) intuition this would have conferred them potential reputational benefits. Importantly, I want to reiterate why in the analyses I did not use the actual presence of cues of accountability, but rather children's subjective perception of such cues. Audience manipulation checks revealed that children who had been assigned to the Audience condition had a suboptimal understanding that they were being held accountable for their decisions. A methodological limitation potentially responsible for this shortcoming is that, in order to increase the believability of the *MegaAttack* game in the Colombian paradigm, I required children to assign a nickname to their own referee-avatar in such a way to resemble common features of PC-games children are usually familiar with. However, this made children in Experiment 3 always nominally identifiable and thus non-anonymous, regardless of whether they were actually subject to cues of accountability (Audience condition) or not (No Audience condition).

Regarding children's punishment-related enjoyment, one of the most interesting findings of my previous work (Experiment 2, Chapter 2) was successfully replicated in the Colombian sample – children's lack of enjoyment in punishing. Moreover, whereas UK children's punishment-related enjoyment (measured after punishment allocation) was on average neutral, Colombian children reported they generally experienced negative rather than neutral affective states. Such negativity was especially apparent in their ratings at the first of the three time points (i.e., before punishing) of Experiment 3. Indeed, Colombian children made a forecasting error about their punishment-related enjoyment in the opposite direction to the one observed by Carlsmith et al. (2008) with US adults. Specifically, Colombian children anticipated punishment to feel worse (instead of better) than how it actually felt during and after punishment allocation. Thus, by adding this valuable cross-cultural dimension to my research I have strengthened my claim that retribution is unlikely to be a primary motive of children's 3PP. Finally, I predicted that inducing children to focus on the outcomes the punishment has on the transgressor while questioning them about their punishment-related enjoyment would make them feel worse. In fact, it was found that question focus (outcome vs no outcome) was not a significant predictor of children's punishment-related enjoyment. Although I did not include any manipulation check to verify whether the wording of the enjoyment question was effective in activating punishment outcome representations, I can speculate that when children are asked about the intentional 3PP action they themselves have carried out, their affective states might be not responsive to variations in the outcome emphasis of such action. Sensitivity to intentions in punishment-related enjoyment, like in judgments of transgression severity and punishment severity decisions, might outweigh sensitivity to outcomes even when the behaviour being evaluated is one's own rather than that of someone else.

Finally, it is important to acknowledge that the sample of Colombian children that was recruited for this experiment was smaller in size than hoped for. Such a small sample size (N = 44) might have prevented the detection of effects when they were in fact present because of lack of statistical power. Moreover, this shortcoming might have also created issues of reliability for the effects that were indeed detected. Therefore, the current findings of Experiment 3 should be regarded as preliminary. In fact, further data collection using the same experimental paradigm but on another Spanish-speaking population is already underway.

In conclusion, the present study has replicated in a small sample of Colombian children of low-middle socio-economic background important findings that had been obtained so far only in WEIRD populations. Specifically, I confirmed children's lack of enjoyment in punishing transgressors of norm violations (Experiment 2, Chapter 2), and the earlier onset of the outcome-to-intention shift in judgements of transgression severity compared to punishment severity decisions (Cushman et al., 2013; Gummerum & Chu, 2014). Whereas these elements of children's morality appear to be universal, others might be subject to a certain degree of cross-cultural variation: culture-specific sensitivity to different moral domains seems to affect the importance children attribute to intentions in judging transgression severity, as well as the amount and developmental pattern of punishment inflicted upon transgressors. Therefore, further studies are needed at the intersection between cognitive anthropology and developmental psychology in order to shed light on the development of intent-based morality from a cross-cultural perspective, thus expanding on pivotal work conducted on adults by scholars as Barrett et al. (2016), McNamara et al. (2019), and Hamilton et al. (1983). This would enable a more fine-grained distinction between universal and culture-specific developmental patterns of judgements of transgression severity, punishment severity decisions and punishment-related enjoyment, ultimately enriching understanding about proximate and evolutionary causes of our socio-moral behaviour.

Chapter 4: Experiment 4

THE UNEXPECTED VALUE OF DETERRENCE IN THIRD-PARTY PUNISHMENT

4.1. Introduction

4.1.1. General introduction

Experiment 4 was built upon the results of Experiments 1-2 (Chapter 2) and planned to be run across three different countries (UK, Colombia and Italy), in parallel with Experiment 3. Differently from Experiment 3, children in Experiment 4 witnessed a wider range of norm transgressions (not limited to issues of unfairness and disloyalty), and had the possibility to restore justice by enacting not one but two types of third-party interventions: punishment of transgressors (i.e., third-party punishment or 3PP) and/or compensation of victims (i.e., third-party compensation or 3PC). This choice was motivated by the fact that, whereas punitive justice has received a great deal of academic attention, compensatory justice still lags behind (Gummerum et al., 2016). Furthermore, very little research testing simultaneously both types of third-party interventions has been conducted from a developmental perspective, and none has adopted a cross-cultural approach. For this reason, in the current Chapter I intended to shed light on the factors influencing on one hand children's compensatory and punitive decisions, and on the other hand children's punishment- and compensation-related affective states. By drawing upon the findings of Experiments 1-2, I narrowed down the list of candidate modulating factors of compensatory and punitive decisions to judgement of transgression severity, type of moral transgression and children's age. Concurrently, the variables I selected as candidate modulating factors of children's affective states were judgement of transgression severity, type of third-party intervention, time passed since the intervention, children's age and endorsement of retribution vs deterrence. With regard to 3PP specifically, I further aimed to deepen the understanding of the justifications and motivations (deterrence vs retribution) leading children to engage in such behaviour. Specifically, I evaluated whether children's 3PP explicit justifications varied across age, cultural background of upbringing, or framing messages about the scope of 3PP children received during the experiment. I also sought to establish which framing message was most in line with children's 3PP implicit motivation.

Besides the theoretical goal of understanding children's responses to transgressions, this study also piloted a new method of conducting experimental research on young children. The specific behaviour lab set-up I utilised for Experiment 4 was an online virtual environment: a Justice System based on the world of *Minecraft*, a globally popular commercial videogame. To test the validity of this method, child participants using this Minecraft Justice System were met either face-to-face or over the internet, through video chat and voice call applications. The results obtained with the two settings were then compared. Thus, in comparison to Experiment 3, Experiment 4 was a testing field to verify the comparability between children's normative reactions across different methodological settings. Since many young children are now familiar with this type of online environment, I believed this new method could ameliorate the practicalities of collecting behavioural data in the field of developmental psychology by greatly expanding recruitment pools away from local geographic areas.

4.1.2. Third-party interventions: Compensation and Punishment

One line of research has put the two types of third-party interventions in direct comparison to establish if people tend to be compensation- or punishment-oriented. Some studies have provided evidence for preference for punishment (Adams & Mullen, 2015; Miller & McCann, 1979; van Prooijen, 2010), others for compensation (Chavez & Bicchieri, 2013; Lotz et al., 2011; Van Doorn, Zeelenberg, & Breugelmans, 2018; Van Doorn, Zeelenberg, Breugelmans, Berger, & Okimoto, 2018). These mixed results might be due to whether the third-party intervention options at participants' disposal were personally costly or not. In the studies showing higher willingness to compensate instead of punishing, both 3PP and 3PC were costly to the participants. In contrast, in studies showing preference for punishment over compensation there were no costs associated to either type of third-party intervention (Van Doorn & Brouwers, 2017).

Of the aforementioned studies, only two included children in their sample. Specifically, Miller and McCann (1979) found that 7- to 12-year old children recommended lower levels of both (non-costly) 3PP and 3PC if the transgressor had already received some punishment and if the transgression was accidental rather than intentional. In case of intentional transgressions with severe consequences for the victim in terms of harm and property destruction, children expressed greater willingness to see the transgressor being punished than the victim being compensated, irrespective of age. Will and colleagues (2013) analysed instead the developmental patterns of costly third-party interventions from 9 to 22 years of age. Participants' investments in compensating victims showed a linear increase with age, while investments in punishing transgressors showed a quadratic trend (decrease between 9 and 16 years of age; increase between 16 and 22 years of age). Besides these two studies, to my knowledge there is only one additional study to have investigated children's relative compensatory and punitive tendencies. Riedl, Jensen, Call and Tomasello (2015) showed that 3-year-old children witnessing instances of theft and unfairness preferred to give back lost resources to the victims (i.e., restoration) over making them inaccessible to the transgressors (i.e., 3PP), thus demonstrating themselves to be victim-oriented.

My proposed research intended to expand the knowledge about children's third-party interventions with the specific aim of simultaneously testing various potential modulating factors of children’s compensatory and punitive choices, especially judgement of transgression severity, moral transgression type and children’s age. I explored whether punishment severity, compensation level and endorsement of compensation vs punishment during justice administration would change as a function of how seriously children judged transgressions. When transgressions were deemed more severe, I expected to observe children increasing their levels of both compensation and punishment, but I did not have strong predictions about which one they would endorse when forced to choose between the two. Children’s decision in the forced-choice task will highlight whether they are more victim- or transgressor-oriented.

I also explored whether the type of moral transgression would affect punishment severity, compensation level and punishment vs compensation endorsement during justice administration. I did not have strong predictions in relation to punishment severity and compensation level. Regarding instead punishment vs compensation endorsement, I predicted children attributing more importance to punishment in instances of harm and property destruction (Miller & McCann, 1979), and to compensation in instances of theft and unfairness/inequity (Riedl et al., 2015). In the latter case it seemed indeed plausible that children would endorse compensation, as there is preliminary evidence that children’s choice of type of third-party intervention is at least partially motivated by the desire to even out the resource unbalances experienced by victims (Experiments 1-2, Chapter 2).

Finally, regarding developmental patterns, I expected to observe a decrease in punishment severity across the age range I chose to focus on (7-11 years), which would be in line with my previous findings (Experiments 1-2, Chapter 2) and with those obtained by Will et al. (2013). However, I did not commit to any specific prediction about the developmental pattern related to compensation levels since Experiment 4 was not economically incentivised as that of Will et al. (2013). Additionally, I tested whether children’s endorsement of punishment vs compensation – both in the forced-choice tasks during and after justice administration – was dependent on age. Since Riedl et al. (2015) demonstrated preference for compensation in 3-year-olds, whereas Miller and McCann's (1979) study showed preference for punishment in 7- to 12-year-olds, I hypothesised there might be a shift towards endorsement of punishment over compensation with increasing age.

4.1.3. Punishment motives: Deterrence vs Retribution

The philosophical literature about the motivational basis of 3PP can be organised around two main theories of justice: Deontological theory and Utilitarian/Consequentialist theory. The best known deontological theory is Kant's theory of retribution (Kant, 1790/1952), according to which punishment is a categorical imperative, whose justification is not rooted in attaining a future benefit but in balancing out a past injustice (backward-looking concern). Under this perspective the transgressor must receive their "just deserts", that is a punishment proportionate to the wrong they have committed.

In contrast, utilitarian theories starting with Bentham's theory of deterrence (Bentham, 1780/1948) conceptualise punishment as a means to achieve societal benefits, i.e. prevention of future transgressions (forward-looking concern). Different utilitarian theories have proposed different ways to actualise this preventive goal: the theory of deterrence aims at deterring the transgressor and/or the general community from engaging in wrongdoing by a threat of punishment; incapacitation theory is focused on depriving the transgressor of the actual ability to pursue new transgressions; rehabilitation theory seeks to re-educate the transgressor in order to change their stance towards wrongdoings.

Regarding psychological research into the matter, investigations making use of a policy-capturing approach (Cooksey, 1996) have provided insight into people's punishment motivations. This technique allows assessment of which information describing transgressions people are more sensitive to when making their punitive decisions. In this way, it has been demonstrated that the severity of participants' punishment recommendations changes depending on retribution-relevant manipulations, but is unaffected by incapacitation- (Darley, Carlsmith, & Robinson, 2000) or deterrence-relevant manipulations, despite participants being willing to invest resources into deterrence of transgressions (Carlsmith et al., 2002). Another method adopted to investigate punishment motivations is the behavioural process tracing task (Jacoby, Jaccard, Kuss, Troutman, & Mazursky, 1987), which aims to identify which information and in which order people seek when making punitive decisions. Before making punishment recommendations people are most likely to first seek retribution-relevant information as it increases their confidence in the appropriateness of their decisions more than incapacitation- and deterrence-relevant information (Carlsmith, 2006; Keller, Oswald, Stucki, & Gollwitzer, 2010).

Of note, a remarkable discrepancy has been noticed between people's actual punitive choices and their explicit justifications. Indeed, people show support for deterrence policies in the abstract but they reject them once seen in operation. This indicates that people fail to anticipate they would perceive as unfair what contradicts the retributive principle of proportionality between transgression and punishment (Carlsmith, 2008). In an additional study, after having read a hypothetical scenario involving a transgression, participants were asked to provide a punishment recommendation for the transgressor and justify their decision. During a semi-structured interview the majority of participants were shown to persist in their original punishment recommendation even when the interviewer pointed out that none of their deterrence- or incapacitation-related justifications were applicable to the specific scenario. These findings not only confirm that participants' punishment decisions were primarily motivated by retribution, but they also suggest that heuristic processes rather than rational deduction may be particularly well suited to explain the retributive motive (Aharoni & Fridlund, 2012).

A more recent study conducted by Crockett, Özdemir and Fehr (2014) analysed actual costly punitive behaviour instead of punishment recommendations. Its experimental conditions were designed to rouse one of two punishment motivations, either deterrence (i.e., open punishment condition) or retribution (i.e., hidden punishment condition). In the open punishment condition transgressors were aware their resource loss was due to third-parties assigning them punishment, while in the hidden punishment condition they were led to believe their loss was due to chance rather than to a third-party's decision. It was observed that third-parties in the hidden punishment condition sanctioned transgressors almost as frequently as in the open punishment condition. This denotes that people are willing to enact costly punishment even when there is no possibility the transgressors could learn from their mistakes and thus be prevented from misbehaving again. Finally, when asked to report their justifications for punishment, third-parties' explanations did not correlate with their behaviour as their endorsement of deterrence motivations far exceeded that of retribution motivations.

In the face of a consistent body of evidence suggesting that adults are actually motivated by retribution despite the utilitarian justifications they provide, it is still unclear whether this mismatch between explicit justifications and implicit punishment motivations is present also in children. The developmental literature on the topic includes interview-based research studies that have shown that US children's rationalisations of punishment may incorporate utilitarian justifications starting from 7-8 years of age (Stern & Peterson, 1999); the same has been observed among older children and early adolescents in Ghana (Twum-Danso Imoh, 2013). The only experimental studies to date about children's expectations regarding punishment functions has been conducted by Bregant, Shaw and Kinzler (2016) and by Yudkin, Van Bavel and Rhodes (2019). In the former study children were shown a scenario depicting a character stealing an important resource from another character. The theft either remained unpunished or was followed by a punishment. Importantly, such punishment was not decided by the children themselves, therefore it was not classifiable as 3PP. By asking children to predict how the thief would behave in the future in the two cases, it was found that children as young as five already believe that the punished thief will be less likely to misbehave again than the unpunished thief, which is evidence of early deterrence reasoning. Instead, in Yudkin et al.'s (2019) study children could decide whether to engage in 3PP by preventing a transgressor (i.e., a harmful peer) from accessing a playing opportunity. Once questioned about the reasons for their punitive decisions, there were some mentions of the desire to see the transgressor change their behaviour and learn a lesson. Importantly, these punishment justifications correlated with children's actual punishment rates. Therefore, it is possible that children's 3PP behaviour is guided by pedagogical considerations and deterrent rather than retributive intents.

Given the lack of research about the (implicit) motivations leading children to engage in 3PP, I drew upon previous framing studies conducted on adults (Feinberg & Willer, 2013, 2015; van Prooijen, 2010) to explore whether the measures of children's punitive tendencies would change depending on the type of punishment frame they had been assigned to, i.e. whether children's role as third-party punishers was framed as serving a deterrent or retributive purpose. The work of van Prooijen (2010) suggested that adults recommend higher sums of money if such transaction is framed as a means to punish transgressors instead of compensating victims. Feinberg and Willer (2013, 2015) demonstrated that political messages are effective at changing people's attitudes when they are framed in such a way to appeal to the targeted individuals' moral values. By applying the same approach, I argue that the punishment frame most in line with children's pre-existing punishment motivation would be also the most effective at increasing their punishment severity and endorsement of punishment over compensation. However, given the lack of prior evidence, I made no prediction as to whether specifically retribution or deterrence framing would trigger higher punitive tendencies. I also made no prediction about the effects of framing conditions on compensation level.

Moreover, in order to deepen the understanding of children's explicit justifications of the use of 3PP, children were also required to endorse deterrence or retribution in a forced-choice task. In this way I could explore whether children's endorsement of deterrence or retribution remains stable across age, culture and framing conditions. By elucidating these aspects I aimed to shed light on how children's punishment justifications develop into adults' more complex crime and justice attitudes. If utilitarian concepts are mere post-hoc rationalisations of punishment decisions actually made under the influence of a retributive impulse, young children would manifest greater support for retribution than for utilitarian ideas because of their immature inhibitory control skills and scarce socialisation to norms of appropriateness. If instead children conceptualise punishment as a means to prevent transgressions – with little or no age, frame and cross-cultural differences – utilitarian reasoning might have deeper roots than currently assumed (Bregant et al., 2016).

4.1.4. Affective states induced by third-party interventions

Although the studies mentioned in the previous section highlight *when* people actually punish (i.e., even when there are no chances to teach a lesson to transgressors), they do not clarify *what* makes punishment satisfactory for those who have enacted it.

Neuroscientific studies have starting clarifying the affective components involved in punishment and compensation. The activation of the striatum, a key area in the brain's reward circuitry, seems to reflect victims' anticipated satisfaction deriving from retaliating against transgressors in a second-party punishment (2PP) paradigm, in which it is the victim rather than an unaffected observer who punishes the transgressors (De Quervain, Fischbacher, Treyer, & Schellhammer, 2004). Moreover, when comparing the neural correlates of punitive decisions in second- and third-party punishers, both categories of people have shown punishment-related activation in the striatum. However, there were important context-dependent quantitative differences: striatal activation was stronger in second- than in third-party punishers (Strobel et al., 2011). Compensation too appears to be intrinsically rewarding. Activation of the striatum – indicative of anticipatory satisfaction – predicts charitable donations to victims of misfortune (Genevsky, Västfjäll, Slovic, & Knutson, 2013; Harbaugh, Mayr, & Burghart, 2007). More recently, the neural activity involved in 3PP and 3PC has been examined comparatively. It has been observed that whereas the striatum is activated in both types of third-party interventions, the specific network being activated in connection with the striatum differs in the two cases (Hu, Strang, & Weber, 2015).

Psychological research about punishment-related affective states has instead produced somewhat mixed results. Interestingly, people confronted with a free rider in a public goods game (a paradigm where 2PP and 3PP are confounded) predict that taking revenge would make them feel better, in line with a retributive motivation to punish. However, once they have enacted punishment, they end up reporting lower mood than their counterparts that did not have the opportunity to punish. This effect is mediated by the fact that, contrary to participants' expectations, punishment causes rumination instead of bringing closure (Carlsmith et al., 2008). Although people – erroneously – anticipate satisfaction from punishing free riders, this does not seem the primary driver of their punishment decisions. Punishment severity is indeed unaffected by mood manipulations, e.g. people punish even when led to believe their mood has been frozen by a bogus drug (Gollwitzer & Bushman, 2012). Given that results suggest that revenge is not always as "sweet" as was previously assumed, researchers started investigating which conditions could make punishment satisfactory. Gollwitzer and colleagues found that people who take revenge for the wrongs suffered are not more satisfied than those who decide not to avenge, possibly indicating that their retribution-driven hedonic expectations have been frustrated. Interestingly though, avengers experience higher levels of satisfaction than non-avengers upon receiving a message from the transgressor that acknowledges that they have understood why they have been punished. These results have been replicated across different operalisations of satisfaction: satisfaction conceptualised as goal fulfilment measured via implicit tests (Gollwitzer & Denzler, 2009) and satisfaction measured via questionnaire rating scales (Gollwitzer et al., 2011). Moreover, when asked to imagine themselves as punishers, people have been shown to accurately forecast actual punishers' emotional experience as they report they would feel more satisfied in receiving the transgressor's acknowledgement feedback than in receiving no such feedback (Funk et al., 2014). By further examining which features of the transgressor's feedback made the punisher's experience satisfactory, it was found that punishers are particularly sensitive to cues of transgressor's change in moral attitude or behaviour, thus supporting a pedagogical interpretation of punishment motivations (Funk et al., 2014). Therefore, this suggests that punishment is neither satisfying (De Quervain et al., 2004; Hu et al., 2015; Strobel et al., 2011) nor dissatisfying per se (Carlsmith et al., 2008), but that its satisfaction level is in fact dependent on whether the transgressor appropriately reacts to the punishment's communicative intent.

Given the gap in knowledge about children's affective states related to third-party interventions, scientific efforts have now begun focussing on the study of the emotional antecedents (Gummerum, López-Pérez, et al., 2019) and emotional consequences of children's 3PP behaviour (Chapters 2-3). I previously demonstrated that both UK children (Experiment 2, Chapter 2) and Colombian children (Experiment 3, Chapter 3) do not enjoy enacting 3PP. On average UK children reported they experienced neither negative nor positive affective states in punishing, while Colombian children's punishment enjoyment was generally negative instead of neutral. Thus, on the basis of my previous studies, I predicted I would find comparable results also in the current study.

Additionally, I made a series of exploratory analyses to investigate the effects of a variety of potential modulating factors of children's enjoyment of third-party interventions. More specifically, I checked whether enjoyment of punishment would change depending on how severely children judged moral transgressions, as well as on whether they endorsed retribution or deterrence as their justification for punishing transgressors. Such analyses were conducted from a developmental perspective in order to examine whether children of different ages would enjoy punishment and compensation to a different degree. No specific predictions were made concerning these research questions.

I also investigated whether enjoyment would vary according to the type of third-party intervention children were enacting, and to the time passed since the intervention. I predicted that compensation would elicit more enjoyment than punishment. Furthermore, based on Carlsmith at al.'s (2008) findings about the negative effect of punishment-induced rumination on adults' affective states, I expected that children's 3PP-related enjoyment would be time-dependent. Although in Experiment 3 I did not observe any significant difference when comparing reports of enjoyment during and after punishment allocation, I considered the possibility that its small sample size might have prevented the detection of an effect when there was in fact one. Conversely, the higher sample size in Experiment 4 gave me enough statistical power to better discriminate the effect of time on 3PP-related enjoyment. Therefore, I predicted enjoyment would be higher immediately after children have enacted punishment towards the transgressors, and then it would decrease once children have had the time to reflect upon their past punishment choices at the end of the experiment. However, I did not formulate any specific prediction for how compensation-related enjoyment would change across time.

4.2. Method

4.2.1. Materials

Eight short videos depicting players' behaviours in *Minecraft* were recorded and embedded into a *Qualtrics* platform questionnaire to create an online Justice System. An offline version was also developed for the purpose of testing at science fairs where the internet connection was not reliable. The system was formatted to resemble an administrative control-panel interface rather than a questionnaire.

The videos, varying in length from 25 to 54 seconds, represented various instances of moral transgressions during Minecraft play: physical harm; property destruction; harm-related false accusation (control); sanctity/authority transgression; theft of resources; property-related trivial accusation (control); disloyalty/inequity in sharing resources; deception/liberty violation (see Appendix C – section γ1.6 for scenario descriptions).

4.2.2. Sample

After exclusions participants were 123 children (*mean age*: 9.83 ± 1.41 years; *age range*: from 7.05 years to 11.97 years; *sex ratio*: 32 females and 91 males) residing in UK, Colombia or Italy. My choice of countries was opportunistic but informed by the Individualism-Collectivism dimension in the Hofstede (Hofstede, 2001) and GLOBE models (House et al., 2004). According to these scales, British culture can be considered individualistic and Colombian culture collectivistic, with Italian culture somewhere between these extremes.

Exclusions from the dataset amounted to 6 children and were differently motivated: 1 child was excluded because of an experimenter's technical mistake; 3 children did not fall into the set age range (i.e., not younger than 7 or older than 11 years); 2 children had difficulties in comprehending experimenter's questions due to lack of experience with playing Minecraft (see script in Appendix C – section γ1).

Participants were tested in one of two alternative settings: either face-to-face (at science fairs or a technology-themed summer camp) or remotely over the internet (via Skype or WhatsApp video or voice calls, depending on the reliability of the internet connection participants had access to from their homes). Data collection lasted from late June 2018 to the beginning of March 2019. The stopping rule was to collect as much data as possible by the set end date.

The Italian sample, consisting of 33 middle-class children, was tested entirely over the internet (nationwide recruitment). The Colombian sample, formed by 23 upper-class children all recruited at the same summer camp of a large city, was tested entirely in a face-to-face setting. The British sample, coming from a mixed socio-demographic background, was tested in both settings: 35 children over the internet (nationwide recruitment) and 32 face-to-face (recruitment at two different science fairs in the same medium-sized English city). The categorisation of the children in terms of their class was not made through detailed socio-economic status assessment, but rather through informal communications with gatekeepers or general impressions about the catchment areas.

4.2.3. Design

I adopted a mixed design in which the between-subject factors were: *Country of residence* (UK; Colombia; Italy)*; Type of experimental setting* (over the internet; face-to-face)*; Type of frame used to describe children's role* (retribution; deterrence; compensation); *Gender* (male; female); *Age* (between 7 and 11 years of age). The within-subject factors were: *Timing of questions* (during or after justice administration); *Type of moral transgression* (8 different scenarios; see Materials – section 4.2.1). Between-subject nuisance factors that were counterbalanced against other between-subject factors were: *Order* *of transgression appearance* (4 possible orders; see Appendix C – Table γ1 for details)*;* *Order of questions* (2 levels: compensation mentioned before punishment and deterrence mentioned before retribution, or compensation mentioned after punishment and deterrence mentioned after retribution; see Appendix C – sections γ1.8, γ1.9, γ1.10 for details).

The dependent variables I measured during justice administration (i.e., repeatedly after each moral transgression) were:

Children's judgement of transgression severity: 6 options, ranging from -5 = "very bad" to 0 = "neither bad nor good".

Levels of punishment of the transgressor: punishment was a ban from the Minecraft server with an 11-point ordinal scale: no ban from the server; 1 hour ban; 2 hours; 6 hours; 12 hours; 1 day; 2 days; 4 days; 1 week; 2 weeks; 4 weeks.

Levels of compensation of the victim: between 0 to 10 Minecraft diamonds.

Levels of punishment-related enjoyment: 11 options where -5 was "very bad"; 0 was "neither bad nor good"; +5 was "very good".

Levels of compensation-related enjoyment: 11 options where -5 was "very bad"; 0 was "neither bad nor good"; +5 was "very good".

Type of third-party intervention endorsed: 2 options, i.e. forced-choice between compensation and punishment.

Frame memorisation (intended as a manipulation check): tested with an open-ended question; children's answers were coded as "punishment undetermined motivation", "punishment retribution", "punishment deterrence" or "compensation".

The dependent variables I measured after justice administration (i.e., after children have finished watching moral transgression scenarios) were: *Levels of punishment-related enjoyment*; *Levels of compensation-related enjoyment; Type of third-party intervention endorsed* (all with the same options as during justice administration); *Type of punishment* *justification endorsed* (2 options, i.e. forced-choice between deterrence and retribution) and *Believability of the game* (2 options, i.e. forced-choice between yes and no).

4.2.4. Procedure

Parents gave consent for their children to participate after having received information about the experiment; an opt-out system was applied in Colombia, opt-in in Italy and the UK. In addition to a specific age range (7-11 years), the other requirement for participation was to have at least a small amount of experience with playing Minecraft in order to understand the dynamics between the players during the experiment. The age range in Experiment 4 was narrower than in Experiments 1-3 because I did not trust children younger than 7 to fully grasp the Minecraft setting.

The procedure began with the researcher explaining that during the experiment they would not be Minecraft players themselves but rather judges helping to test a newly set up Justice System for a Minecraft server called Squidcraft. Participants were told that players on the server experiencing misbehaviours from other players could log their complaints into the Justice System. These complaints along with the chat logs between the players and video renditions of the behaviour in question (see Appendix C – section γ1.6 for details of the complaints and chat logs) would then be shown to a Justice System judge for action to be taken. In reality the complaints and chat logs had been previously written, as well as the videos pre-recorded. This element of deception was revealed to the children once the experiment was completed. Participants' role as judges of the Minecraft justice system was to rate the severity of players' moral transgressions and decide the amount of compensation to allocate to the complainers (i.e., victims) and punishment for the accused (i.e., transgressors). Children did not have to pay any economic cost to enact their third-party decisions. In order to avoid ceiling effects with compensation choices, the experimenter specified that server diamonds were limited and discouraged the children from always giving the maximum number of diamonds.

According to the between-subject condition children were assigned to, the purpose of the Justice System was framed by either emphasising its retributive or deterrent or compensatory functions. This frame was repeated twice, paraphrased in different ways (see Appendix C – sections γ1.2 and γ1.4). The experimenter checked twice whether children had memorised the frame: before trial 1, and then again at the halfway point of the experiment, that is between trial 4 and 5. In both frame manipulation checks, children were asked if they remembered the purpose of the Justice System. The experimenter took note of whether their explanations contained mentions of compensation, deterrence, retribution or of a general punitive motivation with no specific links to retribution or deterrence. When children's answers did not match the assigned frame, the experimenter repeated the frame to them. Recordings were blind double-coded (see Appendix C – section γ2 for double-coding criteria).

After children had responded to the first frame manipulation check, the experimenter presented them with the first complaint of the Justice System, thus starting the justice administration phase of the experiment. Following the reading of the relevant chat log and the viewing of the video, children were asked if they believed the accused player had done what the complaining player said they did. In case of an affirmative answer, children were required to judge the accused player's transgression by rating its severity on a 6-point Likert scale ranging from "very bad" to "neither bad nor good". At this point, children were given the possibility to both punish the transgressor (with a ban from the Minecraft server) and compensate the victim (with server diamonds). The order of punishment and compensation-related questions was counterbalanced across participants.

Right after children decided to enact punishment and/or compensation (i.e., during justice administration), the experimenter asked them to use an 11-point Likert scale ranging from “very bad” to “very good” (whose first 6 points were identical to the scale used to rate judgements of transgression severity) to indicate how they felt in punishing and/or compensating. A substantial difference between Experiment 3 and Experiment 4 should be noted in this context. Whereas in Experiment 3 the enjoyment question during justice administration was asked only once (i.e., once children had punished for the first time), in Experiment 4 the same question was asked multiple times (i.e., every time children had punished a moral transgression).

If children had decided to assign both punishment and compensation, they answered a forced-choice question about whether they considered the former or the latter more important in this specific case (the ordering of the forced-choice question was counterbalanced across participants with regard to whether punishment or compensation was mentioned first).

The procedure from trial 1 was repeated in trials 2 to 4. At the end of trial 4 children were re-asked if they remembered the purpose of the Justice system. Then trials 5 to 8 followed the same procedure as the previous ones. All participants were presented with the same 8 complaints but their order of appearance was counterbalanced across participants.

When all the 8 complaints had been judged (i.e., after justice administration), participants had to answer the final block of questions. Children had to rate on the Likert scale how performing acts of punishment and compensation made them feel (it should be noted that both in Experiments 3 and 4 the enjoyment question after justice administration was asked only once); whether they attributed more importance to punishing transgressors or compensating victims; and whether their main reason for punishing transgressors was for deterrence or retribution. The internal order of these questions was counterbalanced across participants. Finally the experimenter checked whether children believed they had judged misbehaviours that had actually happened on the Minecraft server.

4.2.5. Analysis Strategy and Statistics

The key **dependent variables** of Experiment 4 were the following: *punishment severity* (continuous); *compensation level* (continuous); *punishment vs compensation endorsement during justice administration* (binary) and *after justice administration* (binary); *retribution vs deterrence endorsement* (binary); *punishment-related enjoyment* (continuous); *compensation-related enjoyment* (continuous).

To test my research hypotheses, I adopted linear models implemented using the lme4 package (Version 1.1-21) in the R programming environment (Version 3.5.1, R Core Team, 2018). Depending on the case, I used different types of functions: *lmer* to analyse mixed-effects models of continuous dependent variables; *glmer* to analyse mixed-effects models of binary dependent variables; and *glm* to analyse fixed-effects models of binary dependent variables.

All models included the full range of **independent variables** as follows, distinguished into variables of theoretical interest (details in section 4.3.2) and variables of methodological interest (*setting method*; *believability*; *gender*; *country of residence*). Regarding the nuisance variables, only *question order* but not *transgression order* was included in the models because of stronger theoretical reasons to expect an effect (Condon & DeSteno, 2011). Additionally, *type of moral transgressions* and *children's ID* were included as random factors in models of dependent variables measured during justice administration, in which there were multiple data points per individual.

The significance of fixed and random effects were obtained via ANOVA tests comparing the full model with the effect in question with the model without the effect in question.

The statistic I utilised to calculate effect sizes is based on Nakagawa et al.'s paper (2017). I obtained two main values for each model: $R^2_{GLMM(m)}$, which measures fit of the fixed components of the model, and $R^2_{GLMM(c)}$, which measures fit for fixed and random components together. Effect sizes for individual factors were stated as ΔR^2. Importantly, when I calculated the effect size for a random/fixed factor, ΔR^2 was defined as the reduction in $R^2_{GLMM(c)}/R^2_{GLMM(m)}$ when that factor was removed from the model, but all other factors remained.

4.3. Results

4.3.1. Preliminary analyses on variables of methodological interest and nuisance variable

4.3.1.1. Setting Method

The experimental setting being adopted – whether the experiment was conducted remotely over the internet or face-to-face – did not have any effect on any of the key dependent variables: all *ps* > .1 (see Tables 8 to 14).

4.3.1.2. Believability of the game

In total 88% (95% CI [82%, 94%]) of children believed the events showed in the moral scenarios had actually happened in Minecraft. However, such belief did not exert any effect on the key variables: all *ps* > .1 (see Tables 8 to 14).

4.3.1.3. Question order

The order with which questions were asked during the experimental procedure had a significant effect on children's punishment-related enjoyment (*p* = .026, see Table 13). Specifically, punishment-related enjoyment was higher when the question about punishment was asked first, before the question on compensation enjoyment (M = 2.59, SD = 1.83), than when was asked second (M = 1.96, SD = 2.03). Conversely, question order did not have any effect on all the other key variables: all *ps* > .1 (see Tables 8-12 and 14).

4.3.1.4. Gender

Gender did not exert any effect on any of the key variables (all *ps* > .1, see Tables 8-12 and 14), except for punishment-related enjoyment, *p* = .008 (see Table 13). Specifically, male children enjoyed punishing transgressors (M = 2.53, SD = 1.89) more than female children (M = 1.55, SD = 2.00).

4.3.1.5. Country of residence

Children's country of residence was a significant predictor of **endorsement of punishment vs compensation after justice administration** (*p* = .050, see Table 11). In particular, in the UK 48% (95% CI [36%, 61%]) of children endorsed punishment over compensation; in Colombia 68% (95% CI [47%, 89%]) of children endorsed punishment; in Italy 76% (95% CI [60%, 91%]) of children endorsed punishment. Thus, when forced to choose between punishment and compensation, Italian children proved to be more punitive than British (*p* = .009) but not than Colombian children (*p* = .568); no significant differences were found between Colombian and British children in terms of endorsement of punishment (*p* = .099).

Country of residence was also a significant predictor of **punishment-related enjoyment** (*p* = .039, see Table 13). British children enjoyed punishing (M = 1.78, SD = 1.84) less than Italian (M = 3.11, SD = 1.79, *p* < .001) and Colombian children (M = 2.52, SD = 2.14, *p* = .025); no differences were found between Colombian and Italian children (*p* = .105).

4.3.2. Main analyses on variables of theoretical interest

4.3.2.1. Compensatory and Punitive tendencies

4.3.2.1.1. Variables measured during justice administration

Two of the eight different moral scenarios during the experiment were **controls**: one was a harm-related false accusation; the other a property-related trivial accusation. In the former case, none of the 120 children declared that the accused player had done something bad, χ^2 (1) = 118.01, p < .001. In the latter case, children barely expressed negative judgements related to the transgression severity, M = -0.61, 95% CI [-0.84; -0.37]. For this reason, the two control scenarios were excluded from the statistical analyses. The six remaining scenarios being analysed were: physical harm; property destruction; theft of resources; disloyalty/inequity in sharing resources; deception/liberty violation.

Linear mixed-effects analysis (Table 8) revealed that **punishment severity** did not change across ages and framing conditions, however judgement of transgression severity was a significant predictor. Thus, the more severely children judged the transgressions, the harsher they were in punishing the transgressors, B = -1.07. Furthermore, although transgression type was also a predictor of judgement of transgression severity (χ^2 (1) = 30.73, p < .001, ΔR^2 = .055), transgression type had a significant effect on punishment severity independently of judgement of transgression severity. From a visual representation of the data (Figure 11), it appears that some transgression types (harm) elicited higher 3PP severity and others (theft) lower 3PP severity compared to what would be expected given the judgement of their severity.

***Table 8*. Factors influencing punishment severity during justice administration.**

Factor	χ^2	***p***	ΔR^2
Age	2.41	.121	.013
Gender	0.00	.966	.000
Setting Method	0.01	.917	.000
Frame	2.26	.324	.008
Question Order	0.19	.662	.002
Believability	0.01	.906	.000
Judgement of transgression severity	249.04	< .001 ***	.224
Country of residence	2.37	.305	.005
Transgression type	11.55	.001 ***	.007

* $p \leq .05$.   ** $p \leq .01$.   *** $p \leq .001$.

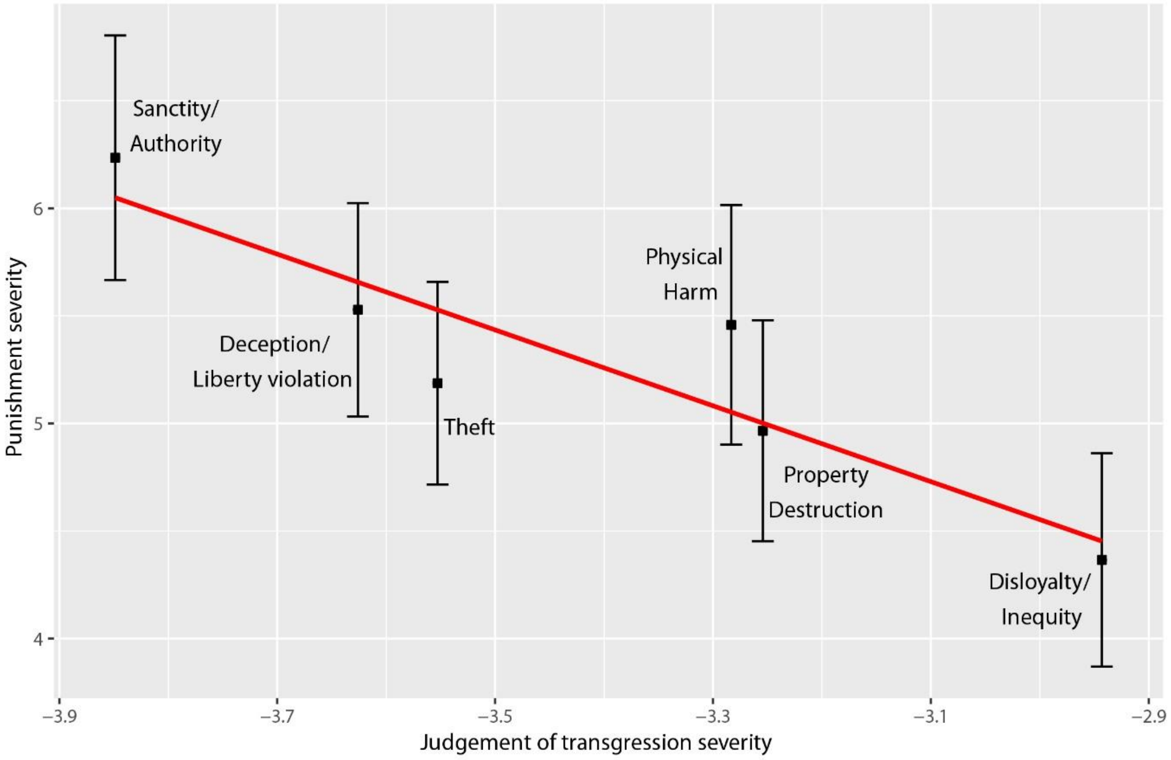

***Figure 11*. Punishment severity in relation to Judgement of transgression severity across transgression types.** 95% CIs are shown for each transgression type; the regression line is based on the means of each transgression type.

As for punishment severity, linear mixed-effects analyses conducted on **compensation level** (Table 9) did not show changes depending on framing conditions. Differently from punishment severity, children's age but not transgression type was a significant predictor of compensation level. Specifically, the younger were the children judging the transgressions, the higher were the compensation levels they enacted towards the victims, B = -0.51. Finally, compensation level was significantly predicted by judgement of transgression severity. Thus, the more severely children judged the transgressions, the more they compensated the victims of such transgressions, B = -0.73.

***Table 9*. Factors influencing compensation level during justice administration.**

Factor	χ^2	***p***	ΔR^2
Age	16.44	< .001 ***	.059
Gender	0.20	.655	.000
Setting Method	1.22	.270	.003
Frame	1.58	.453	.007
Question Order	0.35	.554	.000
Believability	1.76	.185	.005
Judgement of transgression severity	114.21	< .001 ***	.114
Country of residence	5.54	.063	.020
Transgression type	1.79	.181	.005

* $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$.

Generalised linear mixed-effects analyses of children's **endorsement of punishment vs compensation during justice administration** (Table 10) revealed that it was not affected by children's age or frame, whereas judgement of transgression severity and transgression type proved to be significant predictors. Specifically, the more severely children judged the transgressions, the more likely they were to endorse punishment over compensation, B = -0.16. Moreover, it was found that the majority of transgression types did not elicit preferential endorsement either of punishment or compensation (Figure 12). The only two exceptions whereby punishment was clearly the favourite option were for transgressions related to sanctity/authority (73% of children endorsed punishment; 95% CI [64%, 81%]) and liberty/deception (66% of children endorsed punishment; 95% CI [58%, 75%]). Finally, from a visual representation of the data, it is also possible to conclude that instances of theft not only elicited lower 3PP severity (as previously seen in Figure 11) but also lower endorsement of punishment over compensation than what would be expected from the judgement of the severity of such transgression (Figure 12). Further, whereas harm violations elicited higher 3PP severity compared to the expectation deriving from the judgment of transgression severity (as seen in Figure 11), they did not appear to elicit higher endorsement of punishment over compensation (Figure 12).

***Table 10.* Factors influencing punishment vs compensation endorsement during justice administration.**

Factor	χ^2	***p***	**ΔR^2**
Age	2.68	.102	.006
Gender	1.63	.201	.004
Setting Method	0.35	.553	.001
Frame	0.70	.705	.002
Question Order	0.22	.641	.001
Believability	0.62	.430	.001
Judgement of transgression severity	4.50	.034 *	.008
Country of residence	1.63	.444	.005
Transgression type	10.73	.001 **	.032

* $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$.

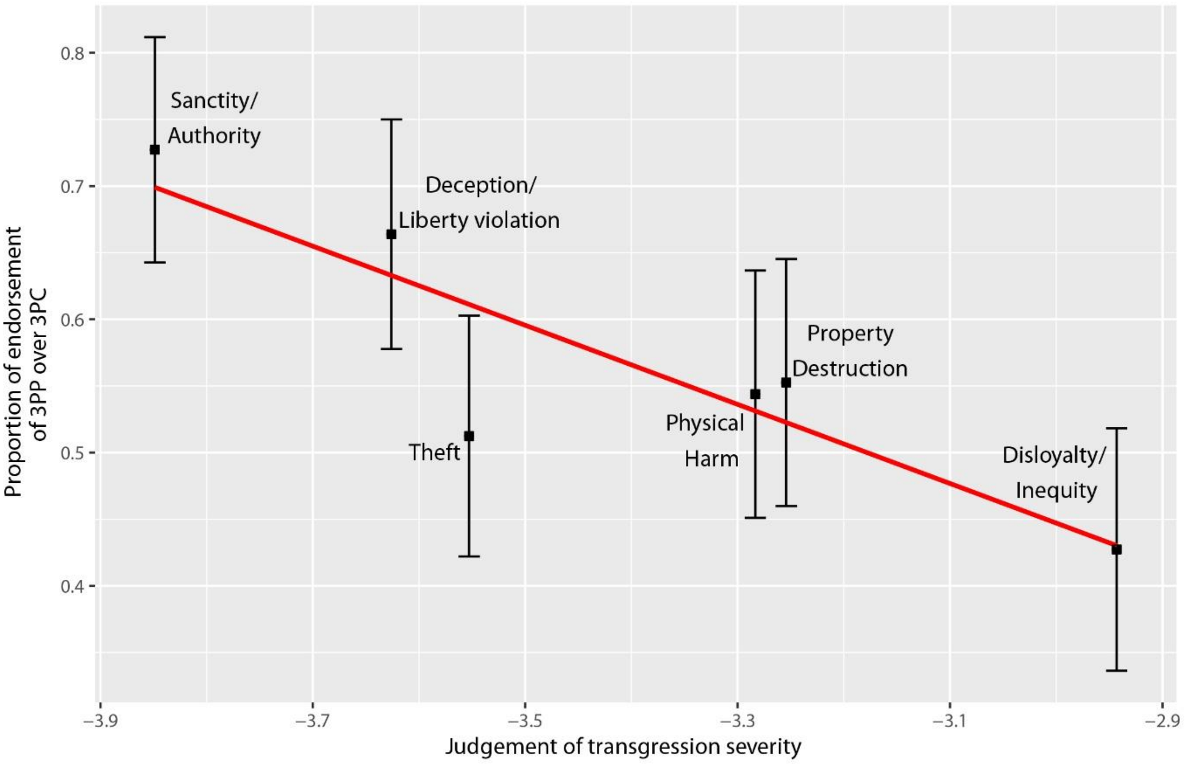

***Figure 12*. Proportion of endorsement of punishment over compensation in relation to judgement of transgression severity across transgression types.** 95% CIs are shown for each transgression type; the regression line is based on the proportions for each transgression type.

4.3.2.1.2. Variables measured after justice administration

After justice administration children's punitive preferences became more clearly detectable. Indeed, 60% (95% CI [51%, 68%]) of children endorsed punishment over compensation in the forced-choice task. **Punishment vs compensation endorsement after justice administration** was not affected by framing condition (Table 11). Age was a significant predictor: the older the children, the more likely they were to endorse punishment over compensation, B = 0.46. The significant effect of children's country of residence is reported in the preliminary analyses as this was not a variable of theoretical interest.

***Table 11*. Modulating factors of punishment vs compensation endorsement after justice administration.**

Factor	χ^2	***p***	**ΔR^2**
Age	7.59	.006 **	.072
Gender	0.61	.436	.006
Setting Method	0.60	.438	.005
Frame	4.31	.116	.041
Question Order	0.06	.810	.000
Believability	2.13	.145	.021
Country of residence	6.00	.050 *	.057

* $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$.

4.3.2.2. Retribution vs Deterrence

4.3.2.2.1. Frame memorisation

In the four models analysed so far (Tables 8-11) it has been shown that framing condition was not a predictor of any DV (all *ps* > .100). Since my frame manipulation did not have any effect on the measures of children's punitive and compensatory tendencies, I turned my attention to our two frame manipulation checks conducted before and half-way through the test trials. I defined "frame memorisation" as the capacity of the children to retain the main information provided by the experimenter during the framing explanation, namely the purpose of the Justice System. Testing whether children's memorisation of the frame changed depending on the type of frame, it was found that the association between framing condition and frame memorisation was significant, both at the first (*p* = .033) and second manipulation checks (*p* = .041), Fisher's exact tests.

Post hoc pairwise comparisons revealed that the degree of memorisation of the deterrence frame was higher than that of the retribution frame, *p* = .034 at the first check and *p* = .040 at the second check. For the degree of memorisation of the compensation frame there was no significant difference from that of retribution (*p* = .220 at the first check and *p* = .344 at the second check) and of deterrence (*p* = .368 at the first check and *p* = .267 at the second check), see Figure 13.

Agreement between two coders on classifying children's explanations about the purpose of the Justice System was investigated for the different coding categories: punishment undetermined motivation; punishment retribution; punishment deterrence; compensation (see Appendix C – section γ2 for double-coding criteria). There was only moderate agreement between the two coders in the classification of punishment undetermined motivation (Cohen's κ = .479), punishment retribution (κ = .576), and compensation (κ = .500). Importantly, there was substantial agreement between the two codes in the classification of punishment deterrence (κ = .794).

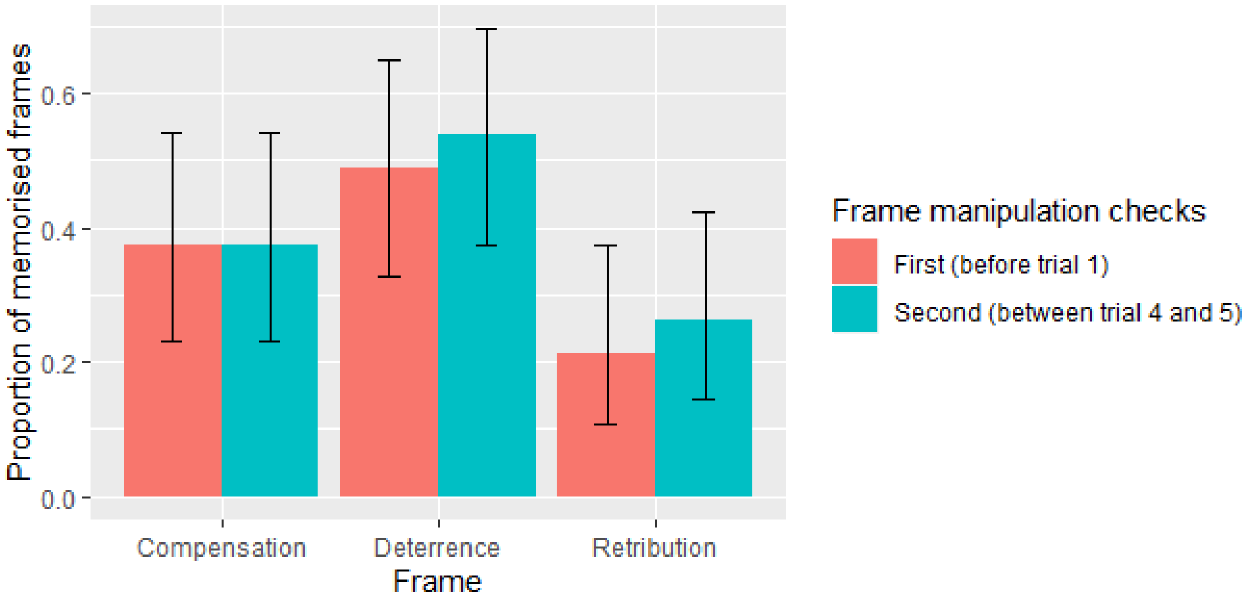

***Figure 13*. Proportion of participants memorising frame across framing conditions.** 95% CIs are shown.

4.3.2.2.2. Punishment justifications

After justice administration 88% (95% CI [80%, 93%]) of children endorsed deterrence over retribution. Generalised linear models (*glm* function for fixed-effects models) revealed that **deterrence vs retribution endorsement after justice administration** was not affected by either children's age or framing condition (Table 12, Figure 14).

***Table 12.* Modulating factors of deterrence vs retribution endorsement after justice administration.**

Factor	χ^2	*p*	ΔR^2
Age	0.18	.668	.002
Gender	0.03	.870	.000
Setting Method	1.13	.287	.011
Frame	4.81	.090	.037
Question Order	0.01	.926	.000
Believability	0.50	.478	.003
Country of residence	1.97	.373	.018

* $p \leq .05$.   ** $p \leq .01$.   *** $p \leq .001$.

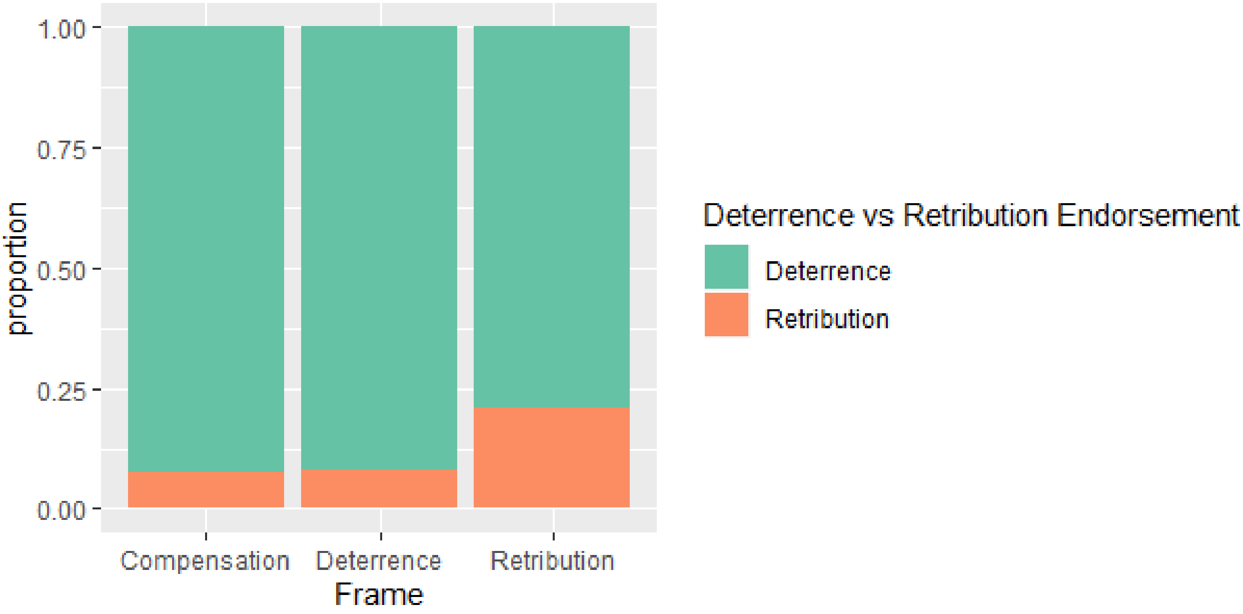

***Figure 14*. Proportion of deterrence vs retribution endorsement across framing conditions.**

4.3.2.3. Enjoyment of third-party interventions

4.3.2.3.1. Comparison between punishment- and compensation-related enjoyment

In order to understand whether enacting punishment or compensation differently affect children's **enjoyment**, I developed a model including as predictors average judgement of transgression severity (across trials), and deterrence vs retribution endorsement in addition to the standard predictors. Average judgement of transgression severity was used as a measure of children's general reactivity to moral transgressions. On the basis of the linear mixed-effects analyses, type of third-party intervention was a significant predictor of children's enjoyment, χ^2 (1) = 76.38, p < .001, ΔR^2 = .093. On average, compensation elicited more enjoyment (M = 3.31, SD = 1.34) than punishment (M = 2.27, SD = 1.96). The effect of the other independent variables on children's enjoyment is reported in the next two sections, broken down for each type of third-party intervention.

4.3.2.3.2. Punishment-related enjoyment

Please note that punishment-related enjoyment during justice administration is an average calculated across trials in order to be directly comparable to punishment-related enjoyment after justice administration (of which there was only one data point per individual). By performing linear mixed-effects analyses (Table 13) I concluded that, whereas frame, children's age and endorsement of deterrence vs retribution did not affect **punishment-related enjoyment**, time was a significant predictor. Enjoyment deriving from punishment was lower when measured after justice administration (M = 2.04, SD = 2.31) than when measured during justice administration (M = 2.50, SD = 1.53). Another significant predictor of punishment-related enjoyment was average judgement of transgression severity. Namely, the more severely children judged transgressions in general, the more they enjoyed enacting punishment, B = -0.65. The significant effects of gender, question order and country of residence are reported in the preliminary analyses as these were not variables of theoretical interest.

***Table 13*. Modulating factors of punishment-related enjoyment.**

Factor	χ^2	*p*	ΔR^2
Time (during or after justice administration)	7.19	.007 **	.015
Age	0.06	.803	.000
Gender	7.08	.008 **	.031
Setting Method	0.03	.864	.000
Frame	0.37	.832	.002
Question Order	4.95	.026 *	.021
Believability	1.27	.260	.005
Average judgement of transgression severity	12.17	< .001 ***	.054
Deterrence vs Retribution endorsement	0.43	.514	.002
Country of residence	6.47	.039 *	.028

* $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$.

4.3.2.3.3. Compensation-related enjoyment

Please note that compensation-related enjoyment during justice administration is an average calculated across trials in order to be directly comparable to compensation-related enjoyment after justice administration (of which there was only one data point per individual). Linear mixed-effects analyses of **compensation-related enjoyment** (Table 14) revealed that time was a significant predictor, meaning that compensation-related enjoyment was higher after justice administration (M = 3.59, SD = 1.44) than during justice administration. (M = 3.05, SD = 1.17). Another significant predictor was average judgement of transgression severity. The more severe were the judgements of transgressions, the more children enjoyed compensating victims, B = -0.69. Finally, framing conditions as well as children's age and endorsement of deterrence vs retribution were not significant predictors of compensation-related enjoyment.

***Table 14.* Modulating factors of compensation-related enjoyment.**

Factor	χ^2	*p*	ΔR^2
Time (during or after justice administration)	19.81	< .001***	.035
Age	2.35	.125	.010
Gender	0.15	.695	.001
Setting Method	2.03	.154	.008
Frame	0.98	.613	.004
Question Order	0.01	.910	.000
Believability	0.01	.909	.000
Average judgement of transgression severity	28.77	< .001***	.132
Deterrence vs Retribution endorsement	0.49	.482	.002
Country of residence	3.96	.138	.016

* $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$.

4.4. Discussion

My research pursued in parallel two main goals: one theoretical (advancing knowledge about children's moral behaviour) and one methodological (validating a new experimental method). Concerning the former point, this research project has produced valuable new information about important but relatively un-investigated topics: the modulating factors of children's third-party interventions, namely punishment of transgressors and compensation of victims; the emotional consequences of enacting punishment and compensation; and the motivations and justifications (deterrence vs retribution) behind children's decisions to inflict punishment from a third-party perspective.

Before moving to an in-depth discussion of our theoretical results, I will briefly dwell upon its methodological findings. Indeed, to my knowledge mine has been the first project in which developmental psychologists have met young children on the internet, by making use of video chat and voice call applications, to test them in an online virtual environment (i.e., a Justice System) positioned within their playground of choice, the world of Minecraft. The lack of any differences in the key variables depending on whether children were tested over the internet or face-to-face, together with the high believability of the experimental procedure, provided evidence that this innovative computer-mediated paradigm has the potential to revolutionise the practicalities of collecting behavioural data in areas with reliable internet connections.

Coming to the theoretical discussion, one of Experiment 4's most interesting results concerns children's explicit justifications for punishment. High levels of endorsement of deterrence over retribution were found, *with no age, frame or cross-cultural differences*, indicating that conceptualising punishment as a means to deter future transgressions is deeply rooted. Explanations linking children's endorsement of deterrence to the development of inhibitory control skills, social desirability bias during the experiment, or socialisation to cultural norms of appropriateness seem therefore unlikely (Bregant et al., 2016). Regarding instead children's implicit punishment motivations it was found that, contrary to my predictions, whether the punishment frame was deterrence or retribution had no effect on children's severity or endorsement of punishment. The failure of our framing manipulation thus suggests that, on average, children do what they think is right, not merely what they are told should be done. Nonetheless, it cannot rule out that this null effect is due to the ineffectiveness of my framing manipulation: the length of the experimenter's explanations about the scope of the justice system might have excessively taxed children's working memory, compromising the possibility to detect *behavioural* differences between children assigned to different punishment frames. It has to be noted, however, that the studies I drawn upon to develop Experiment 4's framing manipulation (Feinberg & Willer, 2013, 2015; van Prooijen, 2010) demonstrated a frame effect on adults' *recommendations* or *attitudes* towards morally-loaded issues rather than on their actual behaviour. Interestingly though, children in Experiment 4 memorised the deterrence frame to a higher degree than the retribution frame, providing evidence that deterrist messages were more easily internalised because they were more aligned with children's pre-existing punishment motivations. It has to be acknowledged that the inter-coder reliability in the classification of children's explanations about the purpose of the Justice System was suboptimal. However, the two coders were in agreement about the higher degree of memorisation of the deterrence frame compared to that of the retribution frame. It remains to be clarified whether the differential degree of internalisation between deterrence and retribution frames was due to the higher frequency at which children are familiarised with adults' deterrist justifications for punishment, or whether children have worked out by themselves that punishment should serve a deterrent scope.

With respect to children's affective states related to third-party interventions, it was found that compensation elicited more enjoyment than punishment, as predicted. This could be due to a "warm glow" effect deriving from the experience of giving to people in need (Andreoni, 1990). Moreover, both compensation- and punishment-related enjoyment were time-dependent but followed different temporal patterns: compensation-related enjoyment increased, while punishment-related enjoyment declined over time. The decrease in punishment-related enjoyment is unlikely to be due to emotional memory extinction (LaBar & Cabeza, 2006) because the same process should have governed compensation-related enjoyment too, which instead showed an increase over time. Whereas the temporal pattern of compensation-related enjoyment might be indicative of children's positive reappraisal of the impact of their action on the victims, the temporal pattern of punishment-related enjoyment is in accordance with Carlsmith et al.'s (2008) finding that enacting punishment causes rumination and thus lowering of mood. It is also possible that the decrease of punishment-related enjoyment over time might also indicate that children experienced a social desirability bias to show regret.

To sum up, children's higher enjoyment of compensation over punishment is not in line with the suggestion of a primarily retributive motivation to punish, which consists of meting out punishment with the expectation of deriving rewarding effects from the transgressors' suffering (Crockett et al., 2014). This conclusion is further supported by children's more proficient internalisation of deterrent over retributive messages, and by their overwhelming endorsement of deterrence irrespective of age, frame and culture. However, an aspect worth highlighting is that in Experiment 4 children's punishment-related affective states were positive, albeit not as positive as compensation-related affective states. This finding is thus in contrast with previous studies that found that children's 3PP-related affective states are neutral (version of the *MegaAttack* paradigm in Experiment 2, Chapter 2) or negative (version of the *MegaAttack* paradigm in Experiment 3, Chapter 3). Although this prevents me from conclusively discarding a retributive explanation for 3PP, I argue that children are neither always nor primarily motivated by retribution. Indeed, the specificities of the different experiments might account for these contrasting results. First of all, in the *MegaAttack* paradigm the allocation of 3PP to the transgressor was more visually and auditorily salient for the participant than in the Minecraft paradigm, probably exerting a greater emotional involvement with the punished transgressor. Moreover, in *MegaAttack* experiments children did not have previous experience with the game, whereas participants in the Minecraft experiment were familiar with Minecraft and enjoyed playing it. Finally and most importantly, in the *MegaAttack* paradigm children could only assign punishment (i.e., 3PP was plausibly felt as a moral obligation), while in the Minecraft paradigm they could both punish transgressors and compensate victims (i.e., contributing to a greater sense of justice being restored).

Relatedly, it was observed that the more severe were the judgements of transgressions, the more children enjoyed both punishing transgressors and compensating victims. As stated before, this could be interpreted as enjoyment deriving from goal fulfilment (i.e., justice restoration) and satisfaction in a job well done. Additionally, analyses revealed that punishment- and compensation-related enjoyment were not affected by children's age. Conversely, it was found that punishment-related enjoyment was influenced by question order, children's gender and country of residence. Specifically, when asked to assign compensation to victims prior to punishment to transgressors, children showed lower punishment-related enjoyment. This could be explained by a compassion effect that carried over from the victim to the punished transgressor (Condon & DeSteno, 2011). Additionally, female (vs male) children and British (vs Italian and Colombian) children reported lower punishment enjoyment, thus opening fascinating questions about the interplay between nature and nurture in the elicitation of moral emotions. Relatedly, previous research that measured 3PP rates instead of 3PP severity found gender differences in punitiveness among Swedish children (Kenward & Östh, 2015), with girls being less punitive than boys. Therefore, it might be that the expectation of lack of punishment-related enjoyment is sufficient to decrease girls' rates of 3PP but not necessarily their levels of 3PP severity.

Furthermore, as expected, the seriousness of a transgression can shape children's third-party interventions. Indeed, the harsher the judgements of transgression severity, the higher the severity of punishment and amount of compensation children assigned respectively to transgressors and victims. Moreover, when forced to endorse either compensation or punishment, children were more likely to express a punitive preference in response to transgressions being judged more severely. This result reveals a transgressor-centred approach to justice restoration, in accordance with other studies in which third-party interventions were not costly to the participants (Adams & Mullen, 2015; Miller & McCann, 1979; van Prooijen, 2010), but in contrast with studies where participants' resources were at stake (Chavez & Bicchieri, 2013; Lotz et al., 2011; Van Doorn, Zeelenberg, & Breugelmans, 2018; Van Doorn, Zeelenberg, Breugelmans, et al., 2018). As Van Doorn and Brouwers (2017) suggested, this could be due to participants acting upon intuition rather than deliberation when they can carry out third-party interventions at no cost to themselves.

Conversely, being required to invest their own resources might induce third-parties to attend more to the victim's needs and to be more careful in choosing the type of intervention conferring them higher reputational benefits.

Contrary to my predictions, children did not preferentially endorse punishment in cases of harm violation and property destruction (Miller & McCann, 1979), and compensation in cases of theft and unfairness/inequity (Riedl et al., 2015). Transgressions in sanctity/authority and liberty/deception were in fact the only contexts eliciting preferential endorsement of punishment. To note, among all the moral norm violations we presented, the scenario depicting a sanctity/authority transgression (i.e., killing a squid, usually allowed but not on this Minecraft server) was the one prompting the harshest judgements as well as the highest levels of punishment severity and endorsement of punishment. Whereas all the other moral norms such as killing and stealing were unconditionally obligatory independent of this Minecraft server, this norm was novel and unique. This is particularly telling, on one hand, of the malleable nature of children's norm learning and, on the other, of the volatility of moral norms on the internet.

Interestingly, while controlling for judgement of transgression severity, type of moral transgression did exert a significant influence on both punishment severity and punishment vs compensation endorsement, but not on compensation level. This suggests that, whereas decisions about compensation levels mostly rely on information about transgression severity, punishment vs compensation endorsement and punishment severity are governed by both information about transgression severity and transgression types. Therefore, even if the type of transgression does not affect the choice of the type of punishment in terms of moral domains (the deep separation model is rejected, Chapter 2), it does affect the severity and endorsement of punishment. This also means that some transgression types prompted higher or lower 3PP severity and endorsement compared to what would be expected on the basis of the judgement of their severity. For example, harm violations evoked higher 3PP severity while instances of theft evoked lower 3PP severity than expected. Theft also elicited higher endorsement of compensation over 3PP than expected.

In order to explain the special "status" of theft transgressions, I refer to the evidence that children use 3PP as an opportunity to directly right the wrong when it concerns resources (Experiment 1, Chapter 2). However, in the current experimental setting 3PP (i.e., giving a time-out from the game to the transgressor) was not suitable for equalising the resource unbalance between victim and transgressor after a theft, so children reacted to this scenario by rather using compensation (i.e., giving diamonds to the victim) to fulfil their equalisation purposes. Instead, the special "status" of the harm violation (i.e., cold-blooded murder of a person) might be explained by the fact that children used 3PP not only to make the transgressor pay for their action, but also to send a message of moral unacceptability for the action itself.

Finally, regarding the investigation of developmental patterns, it was found that compensation levels but not punishment severity decreased with children's increasing age, in contrast with my previous findings (Experiments 1-2, Chapter 2) and with those obtained by Will et al. (2013). Furthermore, as hypothesised on the basis of the studies conducted by Riedl et al. (2015) and Miller and McCann (1979), I observed an increase in the proportion of children endorsing punishment over compensation (after justice administration) with development. This points to the possibility that, although children are willing to punish transgressors from a very early age (Hamlin et al., 2011), attitudes towards this type of third-party intervention are further subject to learning processes. Since compensation is a more socially approved choice than punishment (Patil, Dhaliwal, & Cushman, 2018; Raihani & Bshary, 2015b), I believe it is unlikely that children learn that endorsing punishment over compensation is a way to gain reputational benefits. Instead, I suggest it is more plausible that throughout development children come to better understand the instrumental value of punishment in maintaining social order and communicating messages about the acceptability of specific actions within society (Bregant et al., 2016; Funk et al., 2014). Thus, depending on contextual factors in different societies, I would expect cross-cultural variation in the need to uphold the social contract through punishment. Such speculation is indeed supported by Experiment 4's preliminary evidence that the proportion of children endorsing punishment over compensation (after justice administration) varied across countries (i.e., higher in Italy than in the UK).

Importantly, the samples of children I managed to recruit were not necessarily representative of the respective national populations and this is a limitation. Children tested face-to-face while attending science fairs and technology-themed summer camps came from households characterised by higher education and socio-economic conditions than the national average. In comparison, online testing was more geographically-spread out and potentially more able to reach children of diverse backgrounds. However, only families with a sufficiently good command of video chat and voice call applications could participate in the online experiment (on average, technological literacy was higher among British than Italian parents). Regarding the limitations of my experimental design, I have to acknowledge that I was unable to fully counterbalance the setting method across different countries. Only British children were tested in both settings, whereas Colombian children were tested exclusively face-to-face and Italian children exclusively over the internet because of logistical issues. Having said that, the statistical analyses I conducted always controlled for children's country of residence and setting method, and my aim to test a broad range of children in terms of nationalities was mainly motivated by the desire to maximise the chances of detecting universal patterns of moral behaviour rather than cross-cultural differences.

Future avenues for investigating children's third-party interventions should take advantage of multiple methodologies. Qualitative and possibly longitudinal studies could provide a more detailed insight into the development of children's concepts about punishment justifications. Children might see punishment as a tool to re-establish the norms the society has been built around in order to maintain cooperation and smooth functioning. From interviews with children and their parents and teachers, it would be also possible to discern to what extent children's beliefs about punishment are affected by their familiarity with deterrist justifications in the family and school setting. Questionnaire studies could shed light on personality differences in children endorsing punishment vs compensation and retribution vs deterrence. Areas of interest might include children's assertiveness/dominance, need for order/control, and impulsiveness. Finally, experimental studies could complement the picture by measuring children's affective states (in terms of activation of the brain's reward regions, skin conductance, facial expressions) in three different conditions: when they are given only the opportunity to enact 3PP of transgressors; when they are given only the opportunity to compensate victims; and when they can choose to either punish or compensate. This would allow the scientific community to better disentangle the contribution of the different motives playing a role in children's third-party interventions. Importantly, if a computer game is adopted for testing children's affective states, it is advisable to develop a paradigm that guarantees high believability without requiring children to have previous experience with the game. This would avoid confounding effects between familiarity-induced enjoyment of the game and actual emotional experience of moral decisions. I however encourage future research to test the generalisability of the findings across a variety of settings, by using not only computer-mediated experiments but also puppet shows and experiments conducted with real people.

In conclusion, I have demonstrated that children overwhelmingly reported deterrence as their punishment justification, across different ages and cultures, and even when the system was framed as being for retribution. The deterrence frame was more efficiently internalised by children, probably because more in line with their pre-existing punishment motivations. Moreover, children derived higher enjoyment from compensating victims than from punishing transgressors. Additionally, whereas compensation-related enjoyment increased, punishment-related enjoyment decreased over time. As retribution-motivated people are supposed to adopt punishment because they derive or expect to derive satisfaction from imposing costs upon transgressors, children appeared unlikely to be motivated by retribution. Finally, the more severely children judged the transgressions, the more they endorsed punishment over compensation, revealing a transgressor-centred approach to justice restoration. Taken together, the findings of the current study further theoretical understanding about children's third-party interventions and provide support for the use of a novel, virtual environment as being a viable method for collecting behavioural data with children.

Chapter 5: General Discussion

5.1. Summary of results

The literature in developmental psychology has shown that children intervene as third-party punishers on behalf of victims of harm and fairness transgressions from a young age and even when doing so is economically or socially costly to themselves. However, numerous topics related to children's 3PP specifically, and to the interrelations between 3PP, victim compensation and judgement of transgression severity, remained un-investigated. Of these, the present PhD thesis was aimed at examining the following topics:

- The effects of the kind of moral norm-violations on 3PP, in terms of punishment type, punishment severity and endorsement of punishment over compensation (Experiments 1, 2, 4);
- The emotional experience related to the allocation of 3PP to transgressors and compensation to victims (Experiments 2-4);
- Reputational concerns by evaluating audience effects on 3PP and judgements of transgression severity (Experiments 2-3);
- The effects of intention and outcome information on 3PP and judgements of transgression severity across moral domains (Experiment 3);
- Deterrence vs retribution motives to engage in 3PP (Experiment 4).

In order to research the aforementioned topics, I presented elementary school-aged children across different countries (UK, Colombia and Italy) and socio-economic backgrounds with a variety of moral transgressions, asking them to make their moral evaluations and decisions. The transgressions shown supposedly occurred in an online computer game, either novel (*MegaAttack* game for Experiments 1-3) or already familiar to the children (*Minecraft* for Experiment 4).

In Experiment 1 I measured the type of punishment (fining or banning) children assigned to fairness and loyalty transgressors. My aim was to verify whether children had the tendency to make the punishment fit the crime in terms of moral domains (Graham et al., 2013; Graham et al., 2009). According to the *deep separation model* I had hypothesised, fairness transgressions would motivate an economic type of punishment (fine) whereas loyalty transgressions would motivate a social type of punishment (ban). According instead to the *domain general model*, the type of transgression would have no effect on children's decision about the type of punishment to assign. It turned out that the deep separation model was only partially supported by evidence: children did preferentially assign fines rather than bans to players allocating resources unfairly, but showed no systematic punishment choice preference for disloyal players. Experiment 2 provided potential explanations, although still not conclusive, about this result pattern. Cognitive associative processes might have prompted children to choose a form of punishment employing an element that had a salient role in the transgression scenario. More probably, or in combination, resource equalisation concerns (Shaw & Olson, 2012) might have led children to select the type of punishment allowing them not only to impose a cost on the transgressor but also to equalise as much as possible the resource imbalance between the victim and the transgressor. On that matter, Experiment 4 offered a further demonstration of the importance of inequity aversion in motivating children's moral behaviour. In this experiment, the only type of punishment at children's disposal was banning misbehaving players from the game. However, differently from the other experiments, in Experiment 4 children could also compensate victims by giving them valuable resources. Crucially, when the wrongdoing being judged was a theft, the type of third-party intervention suitable for equalising the resource imbalance between victim and transgressor was compensation rather than 3PP. Therefore, children responded to the theft transgression by endorsing compensation more than 3PP, most likely because the former would guarantee the fulfilment of their equalisation purposes. At the same time, the theft transgression prompted lower 3PP severity compared to what would be expected simply on the basis of the judgement of the transgression severity. Thus, even though the type of transgression did not fully predict the choice of punishment type in terms of moral domains (Experiments 1-2), it proved to be a significant predictor of the severity and endorsement of punishment (Experiment 4).

Regarding the investigation of the emotional experiences associated with justice restoration, I measured children's punishment-related enjoyment in isolation (Experiments 2-3) and in comparison to compensation-related enjoyment (Experiment 4). When children could restore justice only by resorting to 3PP, the rates of punishment were high but children did not enjoy punishing transgressors. Specifically, in Experiment 2 British children's emotional experience of punishing was on average neither positive nor negative. Additionally, the number of children reporting no punishment enjoyment was higher when they were told their 3PP decisions were going to be actually inflicted on transgressors, in comparison to when they were told their decisions were just pretend. In Experiment 3 Colombian children's emotional experience of punishing was negative rather than neutral. In particular, they anticipated punishment to feel worse than how it actually felt during and after punishment allocation. Thus, children's lack of punishment enjoyment in Experiments 2-3 was surprising in the context of the broader literature on the topic: neuroscientific studies have demonstrated that the experience of 3PP is linked to the activation of key areas in the brain's reward circuitry in adults (Hu et al., 2015; Strobel et al., 2011), while psychological studies have indicated that adolescents associate 3PP with positive emotions (Hao et al., 2016). The only studies suggesting that punishers felt worse (Carlsmith et al., 2008) or at least no better than non-punishers (Funk et al., 2014; Gollwitzer & Denzler, 2009; Gollwitzer et al., 2011) were using 2PP paradigms or variations of the public goods game, where 2PP and 3PP are confounded. I speculate that 3PP decisions represent a cognitively demanding task for children, while for adults and gradually for adolescents 3PP choices are easily and intuitively made (see section 5.2 for further details). The age-dependent levels of effort involved in 3PP decision-making might thus be at the origin of the differential enjoyment of the intervention. My findings are also consistent with the fact that children are gradually socialised to believe that wrongdoing deserves punishment. This gradual socialisation into a deservingness belief might causes more enjoyment in adolescents and adults.

Moreover, in order to deepen the knowledge about children's emotional experience of third-party interventions, I compared children's 3PP enjoyment with their compensation enjoyment in Experiment 4. As expected, children enjoyed compensating victims more than punishing transgressors, in line with the so-called "warm glow" effect (Andreoni, 1990). Additionally, compensation enjoyment increased while punishment enjoyment decreased over time. The former temporal pattern could be due to children's positive reappraisal of the impact of their action on the victims; the latter pattern could be induced by rumination about the transgressors' suffering following punishment or by a social desirability bias to show regret. Notably, in Experiment 4 children's emotional experience of punishment was positive, although not as positive as that of compensation. Probably this difference in findings between Experiments 2-3 and Experiment 4 could be explained as follows: in the *MegaAttack* paradigm children could only assign punishment, therefore 3PP was plausibly perceived as a moral duty that ought to be delivered. Conversely, in the *Minecraft* paradigm children could decide to both punish and compensate, thus increasing their sense of goal fulfilment (i.e., justice restoration).

With regard to children's reputational concerns, I wanted to verify whether children would express more severe judgements of transgression severity and inflict more severe 3PP in the perceived presence of an audience. The audience was operationalised in terms of cues of being observed (presence of a commentator and other players observing over the internet, and the attention of the experimenter) in Experiment 2, and in terms of cues of being scrutinised and held personally accountable (presence of an animated mentor watching and judging children's refereeing decisions) in Experiment 3. It was found that in Experiment 2 British children were more judgemental of the transgressions when being observed relative to when they were not being observed, mirroring Bourrat, Baumard & McKay's (2011) finding obtained with adult participants. However, no audience effects were detectable in the levels of 3PP severity. This might suggest that signalling higher moral outrage to an audience is perceived to be associated with more reputational gains than signalling higher punitiveness. Such an explanation makes sense in light of the evidence indicating that third-party punishers can be perceived as aggressive and thus be feared more than liked (Eriksson, Andersson, & Strimling, 2016; Gordon et al., 2014; Patil et al., 2018; Raihani & Bshary, 2015a, 2015b). However, it has to be acknowledged that research about the effects of observability cues on moral behaviour has produced highly mixed results including several failed replications (Bradley et al., 2018; Dear et al., 2019; Kelsey et al., 2018; Northover et al., 2017; Pfattheicher & Keller, 2015; Raihani & Bshary, 2012). Additionally, in Experiment 3 Colombian children showed no differences in the severity of either their judgements or 3PP decisions, irrespective of whether they were being held accountable for their choices or not. A potential explanation for the absence of effects of accountability cues is that people tend to base their moral judgements and punishment decisions on the heuristic that reputation is generally at stake, even when an audience is factually absent (Jordan & Rand, 2019; Tennie, 2012). Consequently, even children in the No Audience condition of Experiment 3 might have felt motivated to increase the harshness of their judgements and punishment decisions with the (non-conscious) intuition this would have conferred them reputational gains.

Experiment 3 was also designed to investigate the so-called “outcome-to-intention shift” in a non-WEIRD sample. In other words, I wanted to clarify when Colombian children’s condemnation of accidental transgressions begins to decrease in favour of an intent-based morality, and whether this occurs in parallel across judgements of transgression severity and 3PP severity decisions. According to the *expression view/capacity model/parallel hypothesis* (Killen et al., 2011; Margoni & Surian, 2016; Zelazo et al., 1996), the developmental changes in children’s explicit moral evaluations are due to cognitive changes (i.e., executive functions and explicit theory-of-mind skills) outside the realm of morality, and thus should simultaneously affect judgements of transgression severity and 3PP severity decisions. According instead to the *emergence view/theory model/constraint hypothesis* (Cushman, 2013; Cushman et al., 2013; Martin et al., 2019), the developmental changes in children’s explicit moral evaluations are primarily the consequence of cognitive changes (i.e., conceptual reorganisation) inside the realm of morality, and thus could have different onsets in judgements of transgression severity and 3PP severity decisions. What I found in my Colombian sample was that the outcome-to-intention shift was already detectable for judgements of transgression severity but not yet for 3PP severity decisions, in accordance with previous findings obtained with British children (Gummerum & Chu, 2014), and supporting the emergence view/theory model/constraint hypothesis. Interestingly, the outcome-to-intention shift in Colombian children’s judgements of transgression severity was moral domain-dependent. Judgements of unfairness were of equal severity between accidental and failed intentional transgressions, whilst judgements of disloyalty were harsher for failed intentional than accidental transgressions. While taking into account that non-WEIRD cultures are especially concerned about binding over individualising moral domains, cultural group selection (Richerson & Boyd, 2005) could be well suited to explain why intentionality is more important within the moral domains emphasised by a specific culture. As negative intentions are a stronger predictor of recidivism than accidents, it makes evolutionary sense that people are watchful about clues indicating someone's intention to disregard the moral norms that their own group cares particularly about (e.g., loyalty norms in collectivistic societies).

Finally, I wanted to shed light on whether children are motivated by deterrence or retribution when they decide to deliver 3PP. It has been theorised that people motivated by deterrence enact 3PP to teach a moral lesson to transgressors with the aim of preventing them from reoffending again. Instead, people motivated by retribution would engage in 3PP for the satisfaction derived from seeing the transgressors suffer (Crockett, Özdemir, & Fehr, 2014). Adults appear to be motivated by retribution, despite the deterrent justifications they provide (Aharoni & Fridlund, 2012; Carlsmith, 2006, 2008; Carlsmith et al., 2002; Keller et al., 2010). With regard to children, it has been found they expect punished transgressors to be less likely to misbehave again compared to unpunished transgressors (Bregant et al., 2016) and they report deterrent justifications when questioned about the reasons for their punitive decisions (Stern & Peterson, 1999; Twum-Danso Imoh, 2013; Yudkin et al., 2019). However, it remained to be clarified whether the mismatch between explicit justifications and implicit punishment motivations is present in children as in adults. Therefore, in Experiment 4 the purpose of the Justice System operated by the children was framed in three alternative ways: retribution; deterrence; or compensation. I measured children's frame memorisation rates, and their endorsement of retribution vs deterrence as punishment justifications. Children were shown to overwhelmingly endorse deterrence over retribution. Since endorsement of deterrence was irrespective of age, country and frame, this finding was unlikely to be the product of children's cognitive development (in terms of inhibition control and forward-looking reasoning), socialisation to cultural norms of appropriateness or social desirability bias during the experiment (Bregant et al., 2016). Furthermore, the deterrence frame was better memorised than the retribution frame, thus providing tentative evidence that deterrent messages were more easily internalised because they were more in line with children's pre-existing punishment motivations (Feinberg & Willer, 2013, 2015; van Prooijen, 2010).

Notably, children's believability rate regarding the presented reality of the games increased across studies: it was 55% in Experiment 1 (where voice-overs were used to comment the *MegaAttack* game); 67% in Experiment 2 (where a live-streamer was shown commenting the *MegaAttack* game); 91% in Experiment 3 (which employed live dialogues between internet players in the *MegaAttack* game); and 88% in Experiment 4 (which consisted in a *Minecraft* administrative control-panel interface). Thus, slight procedural modifications enabled the optimisation of computer-mediated paradigms suitable for face-to-face (Experiments 1-4) or over-the-internet testing sessions (Experiment 4). These computerised methodologies are anticipated to drastically improve the practicalities of collecting behavioural data from children because they can simplify the simultaneous manipulation of numerous variables and, in the specific case of internet-mediated testing, they can also obviate the needs of a local geographical focus and substantial physical resources, causing a reduction in expense.

In conclusion, by adopting innovative computerised paradigms I was able to gather experimental evidence that elementary school-aged children do not yet integrate outcome and intention information in their 3PP decisions as adults do. Children use 3PP as a levelling tool to equalise the imbalance between victims and transgressors, but not demonstrably as a signalling tool to accrue reputational benefits. Importantly, their reported deterrent justifications, better internalisation of deterrence framing messages compared to retribution, and lack of punishment enjoyment or lower enjoyment of punishment compared to compensation, together suggest that children's 3PP motivations are primarily deterrent instead of retributive.

5.2. Theoretical speculations and future avenues for investigation

By applying traditional dual-process models of cognition to 3PP motivations (Kahneman, 2011 for an overview), it has been argued that in adults retribution is driven by automatic and unconscious heuristics (Type I process), while deterrence is governed by rational and conscious deliberation (Type II process) (Aharoni & Fridlund, 2012; Keller et al., 2010). This conclusion had socially problematic implications as it would imply that human beings are endowed with an instinctive and innate propensity to inflict 3PP to cause transgressors to suffer. From this perspective, only effortful cognitive processes would have the potential to override such a destructive impulse in order to bring in preventive and pedagogical considerations. However, dual-process theories have been recently re-elaborated to propose a tripartite view of cognition (Evans & Stanovich, 2013; Ingram & Moreno-Romero, 2019; Railton, 2014; Stanovich, West, & Toplak, 2011). According to this view, Type II process remains a unitary construct, whereas Type I processes include both innate responses (Type 1a process) and learned responses through automatisation (Type 1b process). This conceptualisation has already been provisionally applied to explain emotional responses to moral transgressions (Dys & Malti, 2016; Gummerum, López-Pérez, et al., 2019) and the outcome-to-intention shift during childhood (Ingram & Moreno-Romero, 2019). Thus, at odds with the idea that people's moral decision-making becomes increasingly reliant on controlled processes with age (Gummerum, López-Pérez, et al., 2019), I propose that children's 3PP decisions are mostly based on deliberate (Type II) processes taking into account deterrent goals. Thereafter, the repeated presentation of morally salient stimuli (i.e., moral transgressions) would gradually cause – at some point during adolescence – punitive reactions to become automatised as retributive impulses (Type 1b process).

A way to test the contribution of Type I or Type II processes to 3PP cognition would be to adopt intuition and deliberation manipulations. Specifically, if 3PP in adulthood is governed by an intuitive Type 1b process, the probability or severity of adults' punishment responses would be most strongly affected by intuition manipulations, such as inducing emotional choices. Vice versa, if 3PP in childhood is explained by a deliberate Type II process, the probability or severity of children's punishment responses would be most strongly influenced by deliberation manipulations, such as time pressure and cognitive load. During adolescence there would be a shift in the efficacy of the two kinds of manipulations. Moreover, as suggested by Ingram & Moreno-Romero (2019), another way to test this research hypothesis is to measure people's response times during 3PP decision-making. As it is demonstrated that more automatic choices produce shorter response times than choices made under deliberation (Rubinstein, 2007), response times should decrease with participants' increasing age.

Once the cognitive processes motivating the enactment of 3PP in children are clarified, another line of research to further investigate would be children's expressions of emotions following the observation of moral transgressions from a third-party perspective. Specifically, future studies should aim at clarifying whether children manifest moral domain-specific patterns of facial expressions; whether the type and intensity of their facial expressions predict subsequent punitive interventions; and the extent to which children's facial expressions correlate with their neurophysiological indexes and explicit emotion ratings.

According to Moral Foundations Theory each moral domain is linked to a specific emotion, which is elicited by the violation of the respective moral concern (Haidt & Joseph, 2004, 2007). For example, anger would be triggered by harm violation, and disgust by purity violations. Importantly, patterns of facial movements are considered diagnostic of these prototypical emotions (Ekman, Friesen, & Hager, 2002, but see also Barrett, Adolphs, Marsella, Martinez, & Pollak, 2019). Research on adults has provided tentative evidence that there are both emotion-specific moral domains and emotion-unspecific moral domains (Cannon, Schnall, & White, 2011; Franchin, Geipel, Hadjichristidis, & Surian, 2019; Landmann & Hess, 2018), contrary to the predictions of Moral Foundations Theory. I believe that studies on the development of the association between emotions and moral domains would greatly inform this field by clarifying whether emotions are socially constructed or not.

However, in order to be able to proficiently conduct such research, it would be crucial to rely on a database containing validated stimuli, known to activate specific moral concerns/representations across cultures and ages. This would avoid the inconvenience of researchers designing, for example, an experiment with the intention to be representing an act of disloyalty which is in fact interpreted by participants as an instance of unfairness. I recommend these validation endeavours to be undertaken especially for the use of animated videos, as they are less cognitively taxing for young children than questionnaires or vignettes.

5.3. Practical implications

Increased understanding of moral developmental processes has the potential to be practically as well as theoretically fruitful. The study of children's reactions to moral transgressions can indeed inform intervention studies about the acceptance of the criminal justice system.

It has been demonstrated that when court rulings run counter to lay people's 3PP motivation, this has a detrimental effect on people's trust towards the criminal justice system and on their future adherence to legal codes (Nadler, 2005; Tyler, 1990). Since the cardinal principles of the Western penal system are deterrent and rehabilitative in nature (Beccaria, 1764), Robinson and Darley (1997) have suggested that the criminal justice system should probably be reformed in order that sanctions are devised by taking into higher consideration the offence components that are of utmost relevance for retribution theory. They argued that, by partially rejecting deterrent elements from the legal codes, reformers would help ensure that lay people better recognise the authority of the justice system to guide their own moral behaviour. Paradoxically, this would also cause higher abidance to the law and, in their view, reduction of future crimes. Crucially, this argument is based on the premise that retributive intuitions are hard-wired in human cognition and thus fixed (i.e., governed by Type 1a process). However, as summarised in section 5.1, my evidence suggests a strong deterrent motivation leading children to deliver 3PP against moral transgressors. Situated in the context of the broader literature, this finding implies that 3PP motivations are subject to developmental processes, meaning that there is a shift from deterrence- to retribution-motivated punishment somewhere in the passage from childhood to adulthood. Hence, I tried to argue in section 5.2 that, rather than being innate responses to moral transgressions, retributive attitudes are more likely to be learned responses through automatisation (i.e., driven by Type 1b process). Should my research hypothesis be verified, this would have important implications for intervention studies as learned responses can be reversed by manipulating the modulating factors involved in the learning process. Preventing retributive responses from being internalised and "encapsulated" as automatic during adolescence may be preferable to reforming the criminal justice system in a more retributive direction.

5.4. Conclusions

The experimental work presented in this PhD thesis furthers understanding of 3PP in children, in terms of the factors modulating children's decisions as well as their motivations and emotional experiences of 3PP. Results also suggest that the novel computerised experimental methods, developed and used in the current studies can be effective tools, whose use would encourage and facilitate research in the field. Whilst the presented findings are not without limitation, as discussed in detail in each experimental chapter, overall, results have implications for theoretical accounts of the cognitive processes underlying 3PP, methodological implications for future research approaches and also, potentially, practical implications for the implementation of intervention studies to promote greater acceptance of the criminal justice system.

References

Abdai, J., & Miklósi, Á. (2016). The origin of social evaluation, social eavesdropping, reputation formation, image scoring or what you will. *Frontiers in Psychology, 7*, 1772.

Adams, G. S., & Mullen, E. (2015). Punishing the perpetrator decreases compensation for victims. *Social Psychological and Personality Science, 6*(1), 31-38.

Aharoni, E., & Fridlund, A. J. (2012). Punishment without reason: Isolating retribution in lay punishment of criminal offenders. *Psychology, Public Policy, and Law, 18*(4), 599.

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.

Anderson, J. R., Bucher, B., Kuroshima, H., & Fujita, K. (2016). Evaluation of third-party reciprocity by squirrel monkeys (Saimiri sciureus) and the question of mechanisms. *Animal cognition, 19*(4), 813-818.

Anderson, J. R., Kuroshima, H., Takimoto, A., & Fujita, K. (2013). Third-party social evaluation of humans by monkeys. *Nature Communications, 4*, 1561.

Anderson, J. R., Takimoto, A., Kuroshima, H., & Fujita, K. (2013). Capuchin monkeys judge third-party reciprocity. *Cognition, 127*(1), 140-146.

Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *The Economic Journal, 100*(401), 464-477.

Armsby, R. E. (1971). A reexamination of the development of moral judgments in children. *Child Development*, 1241-1248.

Axelrod, R. (1986). An evolutionary approach to norms. *American Political Science Review, 80*(4), 1095-1111.

Balafoutas, L., Grechenig, K., & Nikiforakis, N. (2014). Third-party punishment and counter-punishment in one-shot interactions. *Economics Letters, 122*(2), 308-310.

Balliet, D., & Van Lange, P. A. (2013). Trust, punishment, and cooperation across 18 societies: A meta-analysis. *Perspectives on Psychological Science, 8*(4), 363-379.

Barclay, P. (2006). Reputational benefits for altruistic punishment. *Evolution and Human Behavior, 27*(5), 325-344.

Barrett, H. C., Bolyanatz, A., Crittenden, A. N., Fessler, D. M., Fitzpatrick, S., Gurven, M., . . . Pisor, A. (2016). Small-scale societies exhibit fundamental variation in the role of intentions in moral judgment. *Proceedings of the National Academy of Sciences, 113*(17), 4688-4693.

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest, 20*(1), 1-68.

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.

Beccaria, C. D. (1764). Dei delitti e delle pene [On Crimes and Punishments]. *Livorno [Leghorn]*.

Behne, T., Carpenter, M., Call, J., & Tomasello, M. (2005). Unwilling versus unable: infants' understanding of intentional action. *Developmental Psychology, 41*(2), 328.

Bentham, J. (1780/1948). A Fragment on Government and an Introduction to the Principles of Morals and Legislation.

Bicchieri, C. (2005). *The grammar of society: The nature and dynamics of social norms*: Cambridge University Press.

Boehm, C. (1987). *Blood revenge: The enactment and management of conflict in Montenegro and other tribal societies*: University of Pennsylvania Press.

Bourrat, P., Baumard, N., & McKay, R. (2011). Surveillance cues enhance moral condemnation. *Evolutionary Psychology, 9*(2), 147470491100900206.

Bowles, S., & Gintis, H. (2004). The evolution of strong reciprocity: cooperation in heterogeneous populations. *Theoretical population biology, 65*(1), 17-28.

Boyd, R., Gintis, H., Bowles, S., & Richerson, P. J. (2003). The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences, 100*(6), 3531-3535.

Boyd, R., & Richerson, P. J. (1992). Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and sociobiology, 13*(3), 171-195.

Bradley, A., Lawrence, C., & Ferguson, E. (2018). Does observability affect prosociality? *Proceedings of the Royal Society B: Biological Sciences, 285*(1875), 20180116.

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.

Bshary, R. (2002). Biting cleaner fish use altruism to deceive image–scoring client reef fish. *Proceedings of the Royal Society of London. Series B: Biological Sciences, 269*(1505), 2087-2093.

Bshary, R., & Grutter, A. S. (2002). Asymmetric cheating opportunities and partner control in a cleaner fish mutualism. *Animal Behaviour, 63*(3), 547-555.

Bshary, R., & Grutter, A. S. (2005). Punishment and partner switching cause cooperative behaviour in a cleaning mutualism. *Biology Letters, 1*(4), 396-399.

Bshary, R., & Grutter, A. S. (2006). Image scoring and cooperation in a cleaner fish mutualism. *Nature, 441*(7096), 975.

Bshary, R., Grutter, A. S., Willener, A. S., & Leimar, O. (2008). Pairs of cooperating cleaner fish provide better service quality than singletons. *Nature, 455*(7215), 964.

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.

Burkart, J. M., Bruegger, R. K., & van Schaik, C. P. (2018). Evolutionary Origins of Morality: Insights from Nonhuman Primates. *Frontiers in Sociology, 3*, 17.

Cannon, P. R., Schnall, S., & White, M. (2011). Transgressions and expressions: Affective facial muscle activity predicts moral judgments. *Social Psychological and Personality Science, 2*(3), 325-331.

Carlsmith, K. M. (2006). The roles of retribution and utility in determining punishment. *Journal of Experimental Social Psychology, 42*(4), 437-451.

Carlsmith, K. M. (2008). On justifying punishment: The discrepancy between words and actions. *Social Justice Research, 21*(2), 119-137.

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.

Carlsmith, K. M., Wilson, T. D., & Gilbert, D. T. (2008). The paradoxical consequences of revenge. *Journal of Personality and Social Psychology, 95*(6), 1316.

Casey, N. (2016). Colombia's Congress Approves Peace Accord With FARC. *The New York Times*. Retrieved from https://www.nytimes.com/2016/11/30/world/americas/colombia-farc-accord-juan-manuel-santos.html

Chakroff, A., Dungan, J., Koster-Hale, J., Brown, A., Saxe, R., & Young, L. (2015). When minds matter for moral judgment: intent information is neurally encoded for harmful but not impure acts. *Social Cognitive and Affective Neuroscience, 11*(3), 476-484.

Chavez, A. K., & Bicchieri, C. (2013). Third-party sanctioning and compensation behavior: Findings from the ultimatum game. *Journal of Economic Psychology, 39*, 268-277.

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.

Choi, Y.-j., & Luo, Y. (2015). 13-month-olds' understanding of social interactions. *Psychological Science, 26*(3), 274-283.

Chudek, M., & Henrich, J. (2011). Culture–gene coevolution, norm-psychology and the emergence of human prosociality. *Trends in Cognitive Sciences, 15*(5), 218-226.

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.

Civai, C., Corradi-Dell'Acqua, C., Gamer, M., & Rumiati, R. I. (2010). Are irrational reactions to unfairness truly emotionally-driven? Dissociated behavioural and emotional responses in the Ultimatum Game task. *Cognition, 114*(1), 89-95.

Clutton-Brock, T., & Parker, G. (1995). Punishment in animal societies. *Nature, 373*(6511), 209.

Clutton-Brock, T., Russell, A., Sharpe, L., & Jordan, N. (2005). 'False feeding'and aggression in meerkat societies. *Animal Behaviour, 69*(6), 1273-1284.

Clutton-Brock, T. H., & Parker, G. A. (1995). Punishment in animal societies. *Nature, 373*(6511), 209.

Condon, P., & DeSteno, D. (2011). Compassion for one reduces punishment for another. *Journal of Experimental Social Psychology, 47*(3), 698-701.

Cooksey, R. W. (1996). *Judgment analysis: Theory, methods, and applications*: Academic Press.

Costanzo, P. R., Coie, J. D., Grumet, J. F., & Farnill, D. (1973). A reexamination of the effects of intent and consequence on children's moral judgments. *Child Development*, 154-161.

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.

Crowley, M. J., Wu, J., Molfese, P. J., & Mayes, L. C. (2010). Social exclusion in middle childhood: rejection events, slow-wave neural activity, and ostracism distress. *Social Neuroscience, 5*(5-6), 483-495.

Curry, O. S., Chesters, M. J., & Van Lissa, C. J. (2019). Mapping morality with a compass: Testing the theory of ‘morality-as-cooperation’with a new questionnaire. *Journal of Research in Personality, 78*, 106-124.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition, 108*(2), 353-380.

Cushman, F. (2013). Action, outcome, and value: A dual-system framework for morality. *Personality and Social Psychology Review, 17*(3), 273-292.

Cushman, F., Sheketoff, R., Wharton, S., & Carey, S. (2013). The development of intent-based moral judgment. *Cognition, 127*(1), 6-21.

D'Ettorre, P., Heinze, J., & Ratnieks, F. L. (2004). Worker policing by egg eating in the ponerine ant Pachycondyla inversa. *Proceedings of the Royal Society of London. Series B: Biological Sciences, 271*(1546), 1427-1434.

Dahl, A., Schuck, R. K., & Campos, J. J. (2013). Do young toddlers act on their social preferences? *Developmental Psychology, 49*(10), 1964.

Daniels, J. P. (2019). Former Farc commanders say they are returning to war despite 2016 peace deal. *The Guardian*. Retrieved from https://www.theguardian.com/world/2019/aug/29/ex-farc-rebels-announce-offensive-despite-peace-deal-colombia-video

Darley, J. M., Carlsmith, K. M., & Robinson, P. H. (2000). Incapacitation and just deserts as motives for punishment. *Law and Human Behavior, 24*(6), 659-683.

De Quervain, D. J., Fischbacher, U., Treyer, V., & Schellhammer, M. (2004). The neural basis of altruistic punishment. *Science, 305*(5688), 1254.

de Waal, F. B. (1982). *Chimpanzee politics: Power and sex among apes*.

de Waal, F. B. (2014). Natural normativity: The 'is' and 'ought'of animal behavior. *Behaviour, 151*(2-3), 185-204.

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.

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.

Dixson, H., & Kenward, B. (*in prep*). The development of third-party punishment in a non-WEIRD sample: ni-Vanuatu children punish similarly to Swedish children.

Dunfield, K. A., & Kuhlmeier, V. A. (2010). Intention-mediated selective helping in infancy. *Psychological Science, 21*(4), 523-527.

Dys, S. P., & Malti, T. (2016). It'sa two-way street: Automatic and controlled processes in children's emotional responses to moral transgressions. *Journal of Experimental Child Psychology, 152*, 31-40.

Ekman, P., Friesen, W., & Hager, J. (2002). Facial Action Coding System: the Manual. Research Nexus, Div. *Network Information Research Corp., Salt Lake City, UT, 1*, 8.

Elster, J. (1989). *The cement of society: A survey of social order*: Cambridge University Press.

Engelmann, D., & Nikiforakis, N. (2015). In the long-run we are all dead: On the benefits of peer punishment in rich environments. *Social Choice and Welfare, 45*(3), 561-577.

Engelmann, J. M., Herrmann, E., & Tomasello, M. (2012). Five-year olds, but not chimpanzees, attempt to manage their reputations. *PLoS One, 7*(10), e48433.

Ericksen, K. P., & Horton, H. (1992). "Blood Feuds": Cross-Cultural Variations in Kin Group Vengeance. *Behavior Science Research, 26*(1-4), 57-85.

Eriksson, K., Andersson, P. A., & Strimling, P. (2016). Moderators of the disapproval of peer punishment. *Group Processes & Intergroup Relations, 19*(2), 152-168.

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

Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on psychological science, 8*(3), 223-241.

Farnill, D. (1974). The effects of social-judgment set on children's use of intent information. *Journal of Personality*.

Fehl, K., Sommerfeld, R. D., Semmann, D., Krambeck, H. J., & Milinski, M. (2012). I dare you to punish me—vendettas in games of cooperation. *PLoS One, 7*(9), e45093.

Fehr, E., Fischbacher, U., & Gächter, S. (2002). Strong reciprocity, human cooperation, and the enforcement of social norms. *Human Nature, 13*(1), 1-25.

Fehr, E., & Gachter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review, 90*(4), 980-994.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature, 415*(6868), 137-140.

Feinberg, M., & Willer, R. (2013). The moral roots of environmental attitudes. *Psychological Science, 24*(1), 56-62.

Feinberg, M., & Willer, R. (2015). From gulf to bridge: when do moral arguments facilitate political influence? *Personality and Social Psychology Bulletin, 41*(12), 1665-1681.

FeldmanHall, O., Mobbs, D., Evans, D., Hiscox, L., Navrady, L., & Dalgleish, T. (2012). What we say and what we do: The relationship between real and hypothetical moral choices. *Cognition, 123*(3), 434-441. doi:10.1016/j.cognition.2012.02.001

Flack, J. C., de Waal, F. B., & Krakauer, D. C. (2005). Social structure, robustness, and policing cost in a cognitively sophisticated species. *The American Naturalist, 165*(5), E126-E139.

Flack, J. C., Girvan, M., de Waal, F. B., & Krakauer, D. C. (2006). Policing stabilizes construction of social niches in primates. *Nature, 439*(7075), 426.

Franchin, L., Geipel, J., Hadjichristidis, C., & Surian, L. (2019). Many moral buttons or just one? Evidence from emotional facial expressions. *Cognition and Emotion, 33*(5), 943-958.

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.

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

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.

Genevsky, A., Västfjäll, D., Slovic, P., & Knutson, B. (2013). Neural underpinnings of the identifiable victim effect: Affect shifts preferences for giving. *Journal of Neuroscience, 33*(43), 17188-17196.

Geraci, A., & Surian, L. (2011). The developmental roots of fairness: Infants' reactions to equal and unequal distributions of resources. *Developmental Science, 14*(5), 1012-1020.

Gintis, H. (2000). Strong reciprocity and human sociality. *Journal of Theoretical Biology, 206*(2), 169-179.

Gollwitzer, M., & Bushman, B. J. (2012). Do victims of injustice punish to improve their mood? *Social Psychological and Personality Science, 3*(5), 572-580.

Gollwitzer, M., & Denzler, M. (2009). What makes revenge sweet: Seeing the offender suffer or delivering a message? *Journal of Experimental Social Psychology, 45*(4), 840-844.

Gollwitzer, M., Meder, M., & Schmitt, M. (2011). What gives victims satisfaction when they seek revenge? *European Journal of Social Psychology, 41*(3), 364-374.

Gordon, D. S., Madden, J. R., & Lea, S. E. (2014). Both loved and feared: third party punishers are viewed as formidable and likeable, but these reputational benefits may only be open to dominant individuals. *PloS One, 9*(10), e110045.

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.

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology, 96*(5), 1029.

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry, 23*(2), 101-124.

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.

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*.

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.

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.

Gächter, S., Renner, E., & Sefton, M. (2008). The long-run benefits of punishment. *Science, 322*(5907), 1510-1510.

Haidt, J., & Joseph, C. (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus, 133*(4), 55-66.

Haidt, J., & Joseph, C. (2007). The moral mind: How five sets of innate intuitions guide the development of many culture-specific virtues, and perhaps even modules. *The Innate Mind, 3*, 367-391.

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.

Hamilton, V. L., Sanders, J., Hosoi, Y., Ishimura, Z., Matsubara, N., Nishimura, H., . . . Tokoro, K. (1983). Universals in judging wrongdoing: Japanese and Americans compared. *American Sociological Review*, 199-211.

Hamilton, W. D. (1964). The genetical evolution of social behaviour. II. *Journal of Theoretical Biology, 7*(1), 17-52.

Hamlin, J. K. (2013). Failed attempts to help and harm: Intention versus outcome in preverbal infants' social evaluations. *Cognition, 128*(3), 451-474.

Hamlin, J. K. (2014). Context-dependent social evaluation in 4.5-month-old human infants: The role of domain-general versus domain-specific processes in the development of social evaluation. *Frontiers in Psychology, 5*, 614.

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature, 450*(7169), 557.

Hamlin, J. K., Wynn, K., & Bloom, P. (2010). Three-month-olds show a negativity bias in their social evaluations. *Developmental Science, 13*(6), 923-929.

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.

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.

Harbaugh, W. T., Mayr, U., & Burghart, D. R. (2007). Neural responses to taxation and voluntary giving reveal motives for charitable donations. *Science, 316*(5831), 1622-1625.

Hauser, M. D. (1992). Costs of deception: cheaters are punished in rhesus monkeys (Macaca mulatta). *Proceedings of the National Academy of Sciences, 89*(24), 12137-12139.

Hauser, O. P., Nowak, M. A., & Rand, D. G. (2014). Punishment does not promote cooperation under exploration dynamics when anti-social punishment is possible. *Journal of Theoretical Biology, 360*, 163-171.

Helwig, C. C., Hildebrandt, C., & Turiel, E. (1995). Children's judgments about psychological harm in social context. *Child Development, 66*(6), 1680-1693.

Helwig, C. C., Zelazo, P. D., & Wilson, M. (2001). Children's judgments of psychological harm in normal and noncanonical situations. *Child Development, 72*(1), 66-81.

Henrich, J., & Boyd, R. (2001). Why people punish defectors: Weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *Journal of Theoretical Biology, 208*(1), 79-89.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2-3), 61-83.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., . . . Ziker, J. (2006). Costly punishment across human societies. *Science, 312*(5781), 1767-1770. doi:10.1126/science.1127333

Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science, 319*(5868), 1362-1367.

Herrmann, E., Keupp, S., Hare, B., Vaish, A., & Tomasello, M. (2013). Direct and indirect reputation formation in nonhuman great apes (Pan paniscus, Pan troglodytes, Gorilla gorilla, Pongo pygmaeus) and human children (Homo sapiens). *Journal of Comparative Psychology, 127*(1), 63.

Hilton, B. C., & Kuhlmeier, V. A. (2018). Intention Attribution and the Development of Moral Evaluation. *Frontiers in Psychology, 9*.

Hofstede, G. (2001). *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations*. Sage publications.

Hollos, M., Leis, P. E., & Turiel, E. (1986). Social reasoning in Ijo children and adolescents in Nigerian communities. *Journal of Cross-Cultural Psychology, 17*(3), 352-374.

Horner, V., & Whiten, A. (2005). Causal knowledge and imitation/emulation switching in chimpanzees (Pan troglodytes) and children (Homo sapiens). *Animal Cognition, 8*(3), 164-181.

House, R. J., Hanges, P. J., Javidan, M., Dorfman, P. W., & Gupta, V. (2004). *Culture, leadership, and organizations: The GLOBE study of 62 societies*. Sage publications.

Hu, Y., Strang, S., & Weber, B. (2015). Helping or punishing strangers: neural correlates of altruistic decisions as third-party and of its relation to empathic concern. *Frontiers in Behavioral Neuroscience, 9*, 24.

Hume, D. (1739/2000). A treatise of human nature: A critical edition. In: by David Fate Norton and Mary J Norton.

Imamoğlu, E. O. (1975). Children's awareness and usage of intention cues. *Child Development*, 39-45.

Ingram, G., & Moreno-Romero, C. O. (2019). Dual-process theories, cognitive decoupling and the outcome-to-intent shift: A developmental perspective on evolutionary ethics. Retrieved from https://psyarxiv.com/dc5rz/

Jacoby, J., Jaccard, J., Kuss, A., Troutman, T., & Mazursky, D. (1987). New directions in behavioral process research: Implications for social psychology. *Journal of Experimental Social Psychology, 23*(2), 146-175.

Janssen, M. A., & Bushman, C. (2008). Evolution of cooperation and altruistic punishment when retaliation is possible. *Journal of Theoretical Biology, 254*(3), 541-545.

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

Jensen, K., Call, J., & Tomasello, M. (2007). Chimpanzees are vengeful but not spiteful. *Proceedings of the National Academy of Sciences, 104*(32), 13046-13050.

Jensen, K., Vaish, A., & Schmidt, M. F. (2014). The emergence of human prosociality: aligning with others through feelings, concerns, and norms. *Frontiers in Psychology, 5*, 822.

Jiang, L.-L., Perc, M., & Szolnoki, A. (2013). If cooperation is likely punish mildly: insights from economic experiments based on the snowdrift game. *PloS One, 8*(5), e64677.

Johnson, C. M., Sullivan, J., Jensen, J., Buck, C., Trexel, J., & St. Leger, J. (2018). Prosocial Predictions by Bottlenose Dolphins (Tursiops spp.) Based on Motion Patterns in Visual Stimuli. *Psychological Science, 29*(9), 1405-1413.

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.

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

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.

Jordan, J. J., & Rand, D. G. (2017). Third-party punishment as a costly signal of high continuation probabilities in repeated games. *Journal of Theoretical Biology, 421*, 189-202.

Jordan, J. J., & Rand, D. G. (2019). 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*.

Kahneman, D. (2011). *Thinking, fast and slow*: Macmillan.

Kant, I. (1790/1952). The critique of judgement (JC Meredith, Trans.). *Oxford: Clarendon Press*, *1969*, 41-90.

Kaplan, J. R. (1978). Fight interference and altruism in rhesus monkeys. *American Journal of Physical Anthropology, 49*(2), 241-249.

Kawai, N., Yasue, M., Banno, T., & Ichinohe, N. (2014). Marmoset monkeys evaluate third-party reciprocity. *Biology Letters, 10*(5), 20140058.

Keller, L. B., Oswald, M. E., Stucki, I., & Gollwitzer, M. (2010). A closer look at an eye for an eye: Laypersons' punishment decisions are primarily driven by retributive motives. *Social Justice Research, 23*(2-3), 99-116.

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.

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.

Kenward, B., & Dahl, M. (2011). Preschoolers distribute scarce resources according to the moral valence of recipients' previous actions. *Developmental Psychology, 47*(4), 1054.

Kenward, B., & Östh, T. (2012). Enactment of third-party punishment by 4-year-olds. *Frontiers in Psychology, 3*, 373.

Kenward, B., & Östh, T. (2015). Five-year-olds punish antisocial adults. *Aggressive Behavior, 41*(5), 413-420.

Kerr, N. L., Hymes, R. W., Anderson, A. B., & Weathers, J. E. (1995). Defendant-juror similarity and mock joror judgments. *Law and Human Behavior, 19*(6), 545-567.

Killen, M., Mulvey, K. L., Richardson, C., Jampol, N., & Woodward, A. (2011). The accidental transgressor: Morally-relevant theory of mind. *Cognition, 119*(2), 197-215. doi:10.1016/j.cognition.2011.01.006

Killen, M., Smetana, J. G., & Smetana, J. (2006). Social–cognitive domain theory: Consistencies and variations in children's moral and social judgments. In *Handbook of Moral Development* (pp. 137-172): Psychology Press.

Kirchkamp, O., & Mill, W. (2018). Conditional Cooperation and the Effect of Punishment. Retrieved from https://www.cesifo.org/DocDL/cesifo1_wp7115.pdf

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.

Krupenye, C., & Hare, B. (2018). Bonobos prefer individuals that hinder others over those that help. *Current Biology, 28*(2), 280-286. e285.

Kummer, H. (1967). Tripartite relations in hamadryas baboons. In “Social Communication among Primates” (SA Altmann, ed.).

Kurland, J. A. (1977). *Kin selection in the Japanese monkey*: Karger Publishers.

Kurzban, R., DeScioli, P., & O'Brien, E. (2007). Audience effects on moralistic punishment. *Evolution and Human Behavior, 28*(2), 75-84.

LaBar, K. S., & Cabeza, R. (2006). Cognitive neuroscience of emotional memory. *Nature Reviews Neuroscience, 7*(1), 54.

Landmann, H., & Hess, U. (2018). Testing moral foundation theory: Are specific moral emotions elicited by specific moral transgressions? *Journal of Moral Education, 47*(1), 34-47.

Lee, Y.-e., Yun, J.-e. E., Kim, E. Y., & Song, H.-j. (2015). The development of infants’ sensitivity to behavioral intentions when inferring others’ social preferences. *PLoS One, 10*(9), e0135588.

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.

Lergetporer, P., Angerer, S., Glätzle-Rützler, D., & Sutter, M. (2014). Third-party punishment increases cooperation in children through (misaligned) expectations and conditional cooperation. *Proceedings of the National Academy of Sciences, 111*(19), 6916-6921. doi:10.1073/pnas.1320451111

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

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.

Lucca, K., Pospisil, J., & Sommerville, J. A. (2018). Fairness informs social decision making in infancy. *PloS One, 13*(2), e0192848.

Lukaszewski, A. W., Simmons, Z. L., Anderson, C., & Roney, J. R. (2016). The role of physical formidability in human social status allocation. *Journal of Personality and Social Psychology, 110*(3), 385.

MacDougall-Shackleton, S. A. (2011). The levels of analysis revisited. *Philosophical Transactions of the Royal Society B: Biological Sciences, 366*(1574), 2076-2085.

Margoni, F., & Surian, L. (2016). Explaining the U-shaped development of intent-based moral judgments. *Frontiers in Psychology, 7*, 219.

Marlowe, F. W., Berbesque, J. C., Barr, A., Barrett, C., Bolyanatz, A., Cardenas, J. C., . . . Henrich, J. (2007). More 'altruistic'punishment in larger societies. *Proceedings of the Royal Society B: Biological Sciences, 275*(1634), 587-592.

Marsh, H. L., Stavropoulos, J., Nienhuis, T., & Legerstee, M. (2010). Six-and 9-month-old infants discriminate between goals despite similar action patterns. *Infancy, 15*(1), 94-106.

Marshall, J., Gollwitzer, A., Wynn, K., & Bloom, P. (2019). The development of corporal third-party punishment. *Cognition, 190*, 221-229.

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.

Martin, J., Buon, M., & Cushman, F. (2019). The effect of cognitive load on intent-based moral judgment. Retrieved from https://psyarxiv.com/em9gx/

Mcauliffe, K., Bogese, M., Chang, L. W., Andrews, C. E., Mayer, T., Faranda, A., . . . Santos, L. R. (2019). Dogs do not prefer helpers in an infant-based social evaluation task. *Frontiers in Psychology, 10*, 591.

McAuliffe, K., & Dunham, Y. (2016). Group bias in cooperative norm enforcement. *Philosophical Transactions of the Royal Society B: Biological Sciences, 371*(1686), 20150073.

McAuliffe, K., Jordan, J. J., & Warneken, F. (2015). Costly third-party punishment in young children. *Cognition, 134*, 1-10.

McGraw, K. M. (1985). Subjective probabilities and moral judgments. *Journal of Experimental Social Psychology, 21*(6), 501-518.

McNamara, R. A., Willard, A. K., Norenzayan, A., & Henrich, J. (2019). Weighing outcome vs. intent across societies: How cultural models of mind shape moral reasoning. *Cognition, 182*, 95-108.

Mendes, N., Steinbeis, N., Bueno-Guerra, N., Call, J., & Singer, T. (2018). Preschool children and chimpanzees incur costs to watch punishment of antisocial others. *Nature Human Behaviour, 2*(1), 45.

Meristo, M., & Surian, L. (2013). Do infants detect indirect reciprocity? *Cognition, 129*(1), 102-113.

Meristo, M., & Surian, L. (2014). Infants distinguish antisocial actions directed towards fair and unfair agents. *PloS One, 9*(10), e110553.

Miller, D. T., & McCann, C. D. (1979). Children's reactions to the perpetrators and victims of injustices. *Child Development*, 861-868.

Monnin, T., & Ratnieks, F. L. (2001). Policing in queenless ponerine ants. *Behavioral Ecology and Sociobiology, 50*(2), 97-108.

Monnin, T., Ratnieks, F. L., Jones, G. R., & Beard, R. (2002). Pretender punishment induced by chemical signalling in a queenless ant. *Nature, 419*(6902), 61.

Mulder, R. A., & Langmore, N. E. (1993). Dominant males punish helpers for temporary defection in Superb Fairy-wrens. *Animal Behaviour*.

Nadler, J. (2005). Flouting the Law, 83 Tex. *L. Rev, 1399*, 1407-1431.

Nakagawa, S., Johnson, P. C., & Schielzeth, H. (2017). The coefficient of determination R 2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface, 14*(134), 20170213.

Nelissen, R. M. (2008). The price you pay: Cost-dependent reputation effects of altruistic punishment. *Evolution and Human Behavior, 29*(4), 242-248.

Nelson, S. A. (1980). Factors influencing young children's use of motives and outcomes as moral criteria. *Child Development*, 823-829.

Nielsen, M., & Haun, D. (2016). Why developmental psychology is incomplete without comparative and cross-cultural perspectives. *Philosophical Transactions of the Royal Society B: Biological Sciences, 371*(1686), 20150071.

Nikiforakis, N. (2008). Punishment and counter-punishment in public good games: Can we really govern ourselves? *Journal of Public Economics, 92*(1-2), 91-112.

Nikiforakis, N. (2010). Feedback, punishment and cooperation in public good experiments. *Games and Economic Behavior, 68*(2), 689-702.

Nikiforakis, N., & Engelmann, D. (2011). Altruistic punishment and the threat of feuds. *Journal of Economic Behavior & Organization, 78*(3), 319-332.

Nobes, G., Panagiotaki, G., & Bartholomew, K. J. (2016). The influence of intention, outcome and question-wording on children’s and adults’ moral judgments. *Cognition, 157*, 190-204.

Nobes, G., Panagiotaki, G., & Pawson, C. (2009). The influence of negligence, intention, and outcome on children’s moral judgments. *Journal of Experimental Child Psychology, 104*(4), 382-397.

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.

Nucci, L. P. (2001). *Education in the moral domain*: Cambridge University Press.

Okimoto, T. G., & Wenzel, M. (2009). Punishment as restoration of group and offender values following a transgression: Value consensus through symbolic labelling and offender reform. *European Journal of Social Psychology, 39*(3), 346-367.

Okimoto, T. G., & Wenzel, M. (2011). Third-party punishment and symbolic intragroup status. *Journal of Experimental Social Psychology, 47*(4), 709-718.

Over, H. (2016). The origins of belonging: social motivation in infants and young children. *Philosophical Transactions of the Royal Society B: Biological Sciences, 371*(1686), 20150072.

Patil, I., Dhaliwal, N., & Cushman, F. (2018). Reputational and cooperative benefits of third-party compensation.

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.

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.

Petit, O., & Thierry, B. (1994). Aggressive and peaceful interventions in conflicts in Tonkean macaques. *Animal Behaviour, 48*(6), 1427-1436.

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.

Piaget, J. (1932/1965). *The Moral Judgment of the Child*. London: Kegan Paul.

Piazza, J., & Bering, J. M. (2008). The effects of perceived anonymity on altruistic punishment. *Evolutionary Psychology, 6*(3), 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.

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.

Raihani, N. J., & Bshary, R. (2015a). The reputation of punishers. *Trends in Ecology & evolution, 30*(2), 98-103.

Raihani, N. J., & Bshary, R. (2015b). Third-party punishers are rewarded, but third-party helpers even more so. *Evolution, 69*(4), 993-1003.

Raihani, N. J., & Bshary, R. (2019). Punishment: one tool, many uses. *Evolutionary Human Sciences, 1*.

Raihani, N. J., Grutter, A. S., & Bshary, R. (2010). Punishers benefit from third-party punishment in fish. *Science, 327*(5962), 171-171.

Raihani, N. J., Pinto, A. I., Grutter, A. S., Wismer, S., & Bshary, R. (2011). Male cleaner wrasses adjust punishment of female partners according to the stakes. *Proceedings of the Royal Society B: Biological Sciences, 279*(1727), 365-370.

Raihani, N. J., Thornton, A., & Bshary, R. (2012). Punishment and cooperation in nature. *Trends in Ecology & Evolution, 27*(5), 288-295.

Railton, P. (2014). The affective dog and its rational tale: Intuition and attunement. *Ethics, 124*(4), 813-859.

Rakoczy, H. (2008). Taking fiction seriously: Young children understand the normative structure of joint pretence games. *Developmental Psychology, 44*(4), 1195.

Rakoczy, H., Hamann, K., Warneken, F., & Tomasello, M. (2010). Bigger knows better: Young children selectively learn rule games from adults rather than from peers. *British Journal of Developmental Psychology, 28*(4), 785-798.

Rakoczy, H., Warneken, F., & Tomasello, M. (2008). The sources of normativity: young children's awareness of the normative structure of games. *Developmental Psychology, 44*(3), 875.

Rakoczy, H., Warneken, F., & Tomasello, M. (2009). Young children's selective learning of rule games from reliable and unreliable models. *Cognitive Development, 24*(1), 61-69.

Rand, D. G., & Nowak, M. A. (2011). The evolution of antisocial punishment in optional public goods games. *Nature Communications, 2*, 434.

Ratnieks, F. L., & Visscher, P. K. (1989). Worker policing in the honeybee. *Nature, 342*(6251), 796-797.

Ratnieks, F. L., & Wenseleers, T. (2005). Policing insect societies. *Science, 307*(5706), 54-56.

Ratnieks, F. L., & Wenseleers, T. (2008). Altruism in insect societies and beyond: voluntary or enforced? *Trends in Ecology & Evolution, 23*(1), 45-52.

Ren, R., Yan, K., Su, Y., Qi, H., Liang, B., Bao, W., & de Waal, F. B. (1991). The reconciliation behavior of golden monkeys (Rhinopithecus roxellanae roxellanae) in small breeding groups. *Primates, 32*(3), 321-327.

Richerson, P. J., & Boyd, R. (2005). *Not by genes alone: How culture transformed human evolution*: University of Chicago press.

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. doi:10.1073/pnas.1203179109

Riedl, K., Jensen, K., Call, J., & Tomasello, M. (2015). Restorative Justice in Children. *Current Biology, 25*(13), 1731-1735. doi: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.

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.

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.

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.

Robinson, P., & Darley, J. (1997). The utility of desert. *Northwestern University Law Review, 91*, 453.

Roos, P., Gelfand, M., Nau, D., & Carr, R. (2014). High strength-of-ties and low mobility enable the evolution of third-party punishment. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1776), 20132661.

Rossano, F., Rakoczy, H., & Tomasello, M. (2011). Young children’s understanding of violations of property rights. *Cognition, 121*(2), 219-227.

RStudio Team. (2015). RStudio: integrated development for R. *RStudio, Inc., Boston, MA*, URL http://www.rstudio.com.

Rubinstein, A. (2007). Instinctive and cognitive reasoning: A study of response times. *The Economic Journal, 117*(523), 1243-1259.

Russell, Y. I., Call, J., & Dunbar, R. I. (2008). Image scoring in great apes. *Behavioural Processes, 78*(1), 108-111.

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

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.

Schmidt, M. F., Rakoczy, H., & Tomasello, M. (2012). Young children enforce social norms selectively depending on the violator's group affiliation. *Cognition, 124*(3), 325-333.

Schmidt, M. F., Rakoczy, H., & Tomasello, M. (2013). Young children understand and defend the entitlements of others. *Journal of Experimental Child Psychology, 116*(4), 930-944.

Schmidt, M. F., & Tomasello, M. (2012). Young children enforce social norms. *Current Directions in Psychological Science, 21*(4), 232-236.

Scola, C., Holvoet, C., Arciszewski, T., & Picard, D. (2015). Further evidence for infants' preference for prosocial over antisocial behaviors. *Infancy, 20*(6), 684-692.

Selten, R. (1967). Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperiments, S. 136–168. *Tübingen: JCB Mohr (Paul Siebeck).*

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.

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

Shweder, R., Much, N., Mahapatra, M., & Park, L. (1997). Divinity and the "big three" explanations of suffering. *Morality and Health, 119*, 119-169.

Siegler, R. S., & Chen, Z. (2008). Differentiation and integration: Guiding principles for analyzing cognitive change. *Developmental Science, 11*(4), 433-448.

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. doi:https://doi.org/10.1016/j.chb.2016.10.019

Smetana, J. G. (1981). Preschool children's conceptions of moral and social rules. *Child Development*, 1333-1336.

Smetana, J. G., Schlagman, N., & Adams, P. W. (1993). Preschool children's judgments about hypothetical and actual transgressions. *Child Development, 64*(1), 202-214.

Smith, C. E., Blake, P. R., & Harris, P. L. (2013). I should but I won't: Why young children endorse norms of fair sharing but do not follow them. *PloS One, 8*(3), e59510.

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

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.

Stanovich, K. E., West, R. F., & Toplak, M. E. (2011). The complexity of developmental predictions from dual process models. *Developmental Review, 31*(2-3), 103-118.

Stern, B. L., & Peterson, L. (1999). Linking wrongdoing and consequence: A developmental analysis of children's punishment orientation. *The Journal of Genetic Psychology, 160*(2), 205-224.

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.

Subiaul, F., Vonk, J., Okamoto-Barth, S., & Barth, J. (2008). Do chimpanzees learn reputation by observation? Evidence from direct and indirect experience with generous and selfish strangers. *Animal Cognition, 11*(4), 611-623.

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

Tasimi, A., & Wynn, K. (2016). Costly rejection of wrongdoers by infants and children. *Cognition, 151*, 76-79.

Tennie, C. (2012). Punishing for your own good: the case of reputation-based cooperation. *Behavioral and Brain Sciences, 35*(1), 40-41.

Tinbergen, N. (1963). On aims and methods of ethology. *Zeitschrift für tierpsychologie, 20*(4), 410-433.

Tisak, M. S., & Jankowski, A. M. (1996). Societal rule evaluations: Adolescent offenders' reasoning about moral, conventional, and personal rules. *Aggressive Behavior: Official Journal of the International Society for Research on Aggression, 22*(3), 195-207.

Tisak, M. S., & Turiel, E. (1988). Variation in seriousness of transgressions and children's moral and conventional concepts. *Developmental Psychology, 24*(3), 352.

Tomasello, M. (2014). The ultra-social animal. *European Journal of Social Psychology, 44*(3), 187-194.

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.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology, 46*(1), 35-57.

Twum-Danso Imoh, A. (2013). Children's perceptions of physical punishment in Ghana and the implications for children's rights. *Childhood, 20*(4), 472-486.

Tyler, T. R. (1990). The social psychology of authority: Why do people obey an order to harm others? Retrieved from https://www.jstor.org/stable/pdf/3053620.pdf

Vaish, A., Carpenter, M., & Tomasello, M. (2010). Young children selectively avoid helping people with harmful intentions. *Child Development, 81*(6), 1661-1669.

Vaish, A., Hepach, R., & Tomasello, M. (2018). The specificity of reciprocity: Young children reciprocate more generously to those who intentionally benefit them. *Journal of Experimental Child Psychology, 167*, 336-353.

Vaish, A., Herrmann, E., Markmann, C., & Tomasello, M. (2016). Preschoolers value those who sanction non-cooperators. *Cognition, 153*, 43-51. doi:http://dx.doi.org/10.1016/j.cognition.2016.04.011

Vaish, A., Missana, M., & Tomasello, M. (2011). Three-year-old children intervene in third-party moral transgressions. *British Journal of Developmental Psychology, 29*(1), 124-130.

Van de Vondervoort, J. W., Aknin, L. B., Kushnir, T., Slevinsky, J., & Hamlin, J. K. (2018). Selectivity in toddlers' behavioral and emotional reactions to prosocial and antisocial others. *Developmental Psychology, 54*(1), 1.

Van de Vondervoort, J. W., & Hamlin, J. K. (2017). Preschoolers' social and moral judgments of third-party helpers and hinderers align with infants' social evaluations. *Journal of Experimental Child Psychology, 164*, 136-151.

Van Doorn, J., & Brouwers, L. (2017). Third-party responses to injustice: a review on the preference for compensation. *Crime Psychology Review, 3*(1), 59-77.

Van Doorn, J., Zeelenberg, M., & Breugelmans, S. M. (2018). An exploration of third parties' preference for compensation over punishment: six experimental demonstrations. *Theory and Decision, 85*(3-4), 333-351.

Van Doorn, J., Zeelenberg, M., Breugelmans, S. M., Berger, S., & Okimoto, T. G. (2018). Prosocial consequences of third-party anger. *Theory and Decision, 84*(4), 585-599.

van Prooijen, J. W. (2010). Retributive versus compensatory justice: Observers' preference for punishing in response to criminal offenses. *European Journal of Social Psychology, 40*(1), 72-85.

Vervafcke, H., De Vries, H., & van Elsacker, L. (2000). Function and distribution of coalitions in captive bonobos (Pan paniscus). *Primates, 41*(3), 249-265.

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.

von Rohr, C. R., Burkart, J. M., & Van Schaik, C. P. (2011). Evolutionary precursors of social norms in chimpanzees: a new approach. *Biology & Philosophy, 26*(1), 1-30.

von Rohr, C. R., Koski, S. E., Burkart, J. M., Caws, C., Fraser, O. N., Ziltener, A., & Van Schaik, C. P. (2012). Impartial third-party interventions in captive chimpanzees: a reflection of community concern. *PLoS One, 7*(3), e32494.

von Rohr, C. R., van Schaik, C. P., Kissling, A., & Burkart, J. M. (2015). Chimpanzees' bystander reactions to infanticide. *Human Nature, 26*(2), 143-160.

Walker-Andrews, A. S., & Kahana-Kalman, R. (1999). The understanding of pretence across the second year of life. *British Journal of Developmental Psychology, 17*(4), 523-536.

Watts, D. P. (1997). Agonistic interventions in wild mountain gorilla groups. *Behaviour, 134*(1-2), 23-57.

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.

Wenseleers, T., Badcock, N., Erven, K., Tofilski, A., Nascimento, F., Hart, A., . . . Ratnieks, F. (2005). A test of worker policing theory in an advanced eusocial wasp, Vespula rufa. *Evolution, 59*(6), 1306-1314.

Will, G.-J., Crone, E. A., van den Bos, W., & Güroğlu, B. (2013). Acting on observed social exclusion: Developmental perspectives on punishment of excluders and compensation of victims. *Developmental Psychology, 49*(12), 2236.

Woo, B. M., Steckler, C. M., Le, D. T., & Hamlin, J. K. (2017). Social evaluation of intentional, truly accidental, and negligently accidental helpers and harmers by 10-month-old infants. *Cognition, 168*, 154-163.

Wyman, E., Rakoczy, H., & Tomasello, M. (2009). Normativity and context in young children's pretend play. *Cognitive Development, 24*(2), 146-155.

Young, L., & Saxe, R. (2011). When ignorance is no excuse: Different roles for intent across moral domains. *Cognition, 120*(2), 202-214.

Young, L., & Tsoi, L. (2013). When mental states matter, when they don't, and what that means for morality. *Social and Personality Psychology Compass, 7*(8), 585-604.

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

Yuill, N. (1984). Young children's coordination of motive and outcome in judgements of satisfaction and morality. *British Journal of Developmental Psychology, 2*(1), 73-81.

Yuill, N., & Perner, J. (1988). Intentionality and knowledge in children's judgments of actor's responsibility and recipient's emotional reaction. *Developmental Psychology, 24*(3), 358.

Zelazo, P. D., Helwig, C. C., & Lau, A. (1996). Intention, act, and outcome in behavioral prediction and moral judgment. *Child Development, 67*(5), 2478-2492.

Zucker, E. L. (1987). Control of intragroup aggression by a captive male orangutan. *Zoo Biology, 6*(3), 219-223.

Appendix A: Supplementary Information related to Chapter 2

Table α1. Counterbalance of experimental conditions across participants in Experiment 1.

Group number	Familiarisation Trial 0	Test Trial 1	Test Trial 2	Test Trial 3	Test Trial 4
1	Refereeing familiarisation	Descriptive norm violation – Fairness norm violation	Descriptive norm violation – Loyalty norm violation	Descriptive norm adherence – Fairness norm violation	Descriptive norm adherence – Loyalty norm violation
2	Refereeing familiarisation	Descriptive norm adherence – Fairness norm violation	Descriptive norm adherence – Loyalty norm violation	Descriptive norm violation – Fairness norm violation	Descriptive norm violation – Loyalty norm violation
3	Refereeing familiarisation	Descriptive norm violation – Loyalty norm violation	Descriptive norm violation – Fairness norm violation	Descriptive norm adherence – Loyalty norm violation	Descriptive norm adherence – Fairness norm violation
4	Refereeing familiarisation	Descriptive norm adherence – Loyalty norm violation	Descriptive norm adherence – Fairness norm violation	Descriptive norm violation – Loyalty norm violation	Descriptive norm violation – Fairness norm violation

Table α2. **Counterbalance of experimental conditions across participants in Experiment 2.**

Group number	Familiarisation Trial 0	Test Trial 1	Test Trial 2	Test Trial 3	Test Trial 4
1	Refereeing familiarisation	Fairness norm violation – No Audience	Loyalty norm violation – Audience	Fairness norm violation – Audience	Loyalty norm violation – No Audience
2	Refereeing familiarisation	Loyalty norm violation – Audience	Fairness norm violation – No Audience	Loyalty norm violation – No Audience	Fairness norm violation – Audience
3	Refereeing familiarisation	Fairness norm violation – Audience	Loyalty norm violation – No Audience	Fairness norm violation – No Audience	Loyalty norm violation – Audience
4	Refereeing familiarisation	Loyalty norm violation – No Audience	Fairness norm violation – Audience	Loyalty norm violation – Audience	Fairness norm violation – No Audience

α1. Script of Experiment 1

α1.1. CONSENT PHASE

Following initial greeting, verbal recap of salient points from the information sheet.

<Start Audio Recording>

- Experimenter to the child: "*My boss and I have made this new game you can play both locally or on the internet. In this game you can be either a player or a referee. When you are a player I will be your team member, and we will play together against robot enemies, without being connected to the internet. When you are a referee, other children will connect to this game via the internet to be players. I've noted that sometimes in this game players behave badly towards their team-member. And your job as a referee will be to tell me if the players have behaved badly or not, a bit like in football games but in space!*"

- Consent phrasing: "*It takes about 15 minutes to be in the study. Your parent is happy for you to take part in our project, but remember, you don't have to be in it if you don't want to. So you can stop playing with me if you want, even if we start. If you have any questions, now you can ask them all.* [Pause to answer the questions]. *Alright! Does all this sound OK? You want to do this now?"* [Pause so verbal consent can be recorded]".
- Experimenter explains the keys before starting the game. "*Now let's start with being a player! You just need to remember four keys: the up arrow key is to move forward; the right arrow key is to rotate towards the right; the left arrow key is to rotate towards the left; the space bar is to shoot*"

<u>α1.2. PLAYING FAMILIARISATION</u> (to clarify moral norms)

The child is introduced to every element of the game screen by the experimenter:

- **Gems**. *"This is the game arena in which we are going to play. Do you see these purple gems? We must collect them. All the gems we collect go into this storage at the top, all together because we belong to the same team. And the more gems we collect, the more weapons we can buy later on if we play for a long time."*
- **Enemies**. *"While collecting the gems, we must also defend ourselves or the team-member from the attacks of the enemies – these green rockets – by shooting at them"*
- **Shield**. *"Do you know what this coloured circle around you is? This is the shield that protects you from the enemies' attacks. But remember that it won't last forever so try not to get killed."*
- **Time**. *"And this bar at the top shows that the time is passing."*

The child and the experimenter play together for 4 game bouts <u>(order of events remains the same across participants).</u>

- **1° Game bout**:
    - The experimenter teaches the child how to use the controls.
    - The child plays with the experimenter.
    - The experimenter does not trigger the mega-attack targeting the child player.
    - At the end of the trial there is no gem distribution.
- **2° Game bout**:
    - The child and the experimenter play together.
    - The experimenter does not trigger the mega-attack targeting the child player.
    - At the end of the trial the child has to distribute the gems between themselves and the experimenter. The experimenter makes no comment on the division chosen by the child.
- **3° Game bout**:
    - The child and the experimenter play together.
    - The experimenter does not trigger the mega-attack targeting the child player.
    - **Fairness norm illustration**. At the end of the trial the experimenter distributes the gems between herself and the child; the experimenter always chooses a fair division.
- **4° Game bout**:
    - The child and the experimenter play together.
    - **Loyalty norm illustration**. Making sure to remain unnoticed, the experimenter triggers the mega-attack by pressing X on the keyboard. The experimenter explains: "*this mega-attack is against you, you are not going to survive if you don't get help. But don't worry, I am going to help you with my secret weapon!*".

- ° At the end of the trial the experimenter has to distribute the gems between herself and the child. Just before deciding for a fair split, the experimenter pauses to say: "*After this, we look on the internet for a game to referee. Sometimes on the internet, you hear the comments being made live by a kid who likes to practice their youtubing skills and that sort of thing*".

<u>α1.3. REFEREEING FAMILIARISATION</u> (players do not violate any moral or descriptive norm)

The experimenter speaks during the game to explain the main features to the child.

- **Referee presentation stage**. *"Now it's time for you to referee the game! Observe carefully how the players behave because I will ask you some questions about their behaviour. Look, there's the referee at the bottom. That's you!"*
- **Audience presentation stage**. "*Look, at the edges of the arena there are plenty of internet players. They all want to play, indeed everyone is ready to enter the arena with their blue/red shield* [according to the descriptive norm]*...but only two at a time can play; the others will have to wait and look at the game in the meanwhile.*"
- **Shield choice stage/Descriptive norm establishment**. "*This time the two players are* [name of the 1° player] *and* [name of the 2° player]. *Now we are seeing them choosing their shield. As usual, both players have chosen a blue/red shield* [according to the descriptive norm] *like all the other players in the audience. Now that they have made their choice, they are ready to teleport into the arena to start the game!*"
- **Gem collection stage [players abide by Loyalty norm]**. "*Wow, there a lot of gems that players can collect.* [While players are collecting gems]…*The players are making a really good job, look at how many gems are now in the communal store! The players have been hit a bit by the enemies, but after all, the team is managing to get by without getting hurt too much… Very good, at the end no player has died!*"

- **Gems distribution stage [players abide by Fairness norm]**. "*Now it's time to split the gems which were collected before. These points are very important because they can allow the players to level up and get better weapons. The* [name of the 1° player] *is deciding how to split the points.* [The 1° player splits the points equally]*...and it was a fair split!"*
- **Judgement stage**. "*Remember that now you are the referee, it's your job to decide if they did anything wrong".*
  - [By pointing at the player on the right] "*Did this player do anything wrong?".* The child is expected to answer no, so the experimenter proceeds by saying: *"Now press N* [on the laptop's keyboard] *to go to the next player then".*
  - [By pointing at the player on the left] *"Did this player do anything wrong?".* Also in this case the child is expected to answer NO.
  - If the child answers YES to one or both questions, the experimenter will follow the script for the test trials (see below).

<u>α1.4. TEST TRIALS</u> (players are shown violating moral and/or descriptive norms; order of players' transgressions varies across participants)

The experimenter remains quiet during the recorded commentary.

- **Judgement stage (i.e. children's judgement of transgression severity)**
  - At the end of the scenario, for each player in turn, the experimenter asks the child: "*Did this player do anything wrong?*"
  - If the child answers NO, the experimenter says: "*Press N to go to the next player then*".
  - If the child answers YES, the experimenter asks the two following questions:

- "*What did this player do wrong?*". If the children have a false memory, the experimenter is allowed to correct it, so that she does not accept their claim about a wrong behaviour that did not happen, but she does accept any claim that something was wrong, if it actually happened.
- "*And how bad was the player's behaviour?".* The experimenter shows a 5-point smiley face scale on a paper (Figure α1) and asks the child to point at the face of their choice by saying: "*Choose a face on this scale, where this* [pointing at the face on the right] *is just a little bad, while this* [pointing at the face on the left] *is super bad*".

***Figure α1*. Judgement of transgression severity scale**: 5 points, ranging from 1 = "just a little bad" (on the right) to 5 = "super bad" (on the left).

- **Punishment stage**
  - Experimenter to the child: "*Now press P* [on the laptop's keyboard] *so you can give the mean player a penalty*". At this point the child is required to decide the type and severity of punishment.
  - **Type of punishment**

    Experimenter to the child: "*Now you can give a time-out from the game to the mean player* [experimenter pointing at the icon on the left side of the laptop's screen, see Figure α2] *– so that they wouldn't be allowed to play for a while – or you can take away some of their gems* [experimenter pointing at the icon on the right side of the laptop's screen, see Figure α2]. *Which kind of penalty do* *you want to give the mean player? Move the referee with the arrow keys and then press G when you have decided the type of penalty for the mean player*".

Time out
Lose gems

Figure a2. **Types of punishment**: on the left social punishment (banning from the game); on the right economic punishment (loss of gems).

- **Severity of punishment**
    - If the child chooses **Time out**, the experimenter asks: "*How long do you want the time out to be?*". The child is given six options on a scale shown on the laptop's screen: **0 minutes**; **1 minute**; **5 minutes**; **20 minutes**; **1 hour**; **1 day**.
    - If the child chooses **Lose Gems**, the experimenter asks: "*How many gems do you want the mean player to lose?*". The child is given six options on a scale shown on the laptop's screen: **0 gems**; **2 gems**; **5 gems**; **10 gems**; **50 gems**; **100 gems**.
    - In both cases, the child has to move the referee by using the arrow keys and then press G when they have decided the severity of the punishment.

α1.5. SATISFACTION RATING OF THE ACTIVITY

NB: the following question was asked for advertisement rather than scientific purposes; in order to recruit more children we needed data to support our claim that the activity was enjoyable.

After all the test trials, the experimenter assesses the child's enjoyment of the activity by asking: "*Now that you've made your choices, can I ask you if you have liked being a referee in this game?".* The experimenter shows a 5-point smiley face scale on a paper (Figure α3) and asks the child to circle the face of their choice by saying: *"Can you choose a face on this scale?*".

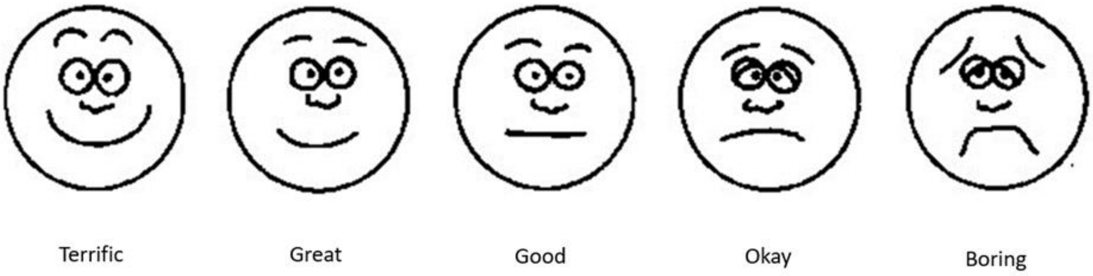

***Figure α3*. Judgement of activity scale**: 5 points, ranging from 1 = "Boring" to 5 = "Terrific".

α1.6. BELIEFS ABOUT PUNISHMENT TYPES

Experimenter to the child: "*As a referee, at the end of each game you could decide whether to give a time-out or take away gems from the mean player. In general, which do you think was worse for the mean player? Receiving a time-out from you or having their gems taken away? Or is it the same?*".

α1.7. REALITY CHECK

- Experimenter to the child: "*Do you think you really refereed games with internet players now?*".
- If the child answers NO, the experiment is considered completed.

- If the child answers YES, then the experimenter further asks: "*Did you really give the internet players time-outs and/or take away gems?*" NB – data was not analysed.
- After the child has replied to this final question, the experiment is considered completed.

<Stop Audio Recording>

α2. Description of video clip with Experiment 1 paradigm

The voice-over comments that the descriptive norm is for the player-avatars to choose red shields, as shown by the audience of avatars outside the game arena. The players teleported into the game arena are Fox and Panda. Player Panda abides by the descriptive norm by appearing with a red shield, while player Fox commits a descriptive norm violation in choosing a blue shield. After each player has collected several gems, player Panda gets surrounded and attacked by computer-controlled enemies in a mega-attack. Player Fox makes a loyalty transgression as they do not come to the aid of the team-mate, whose space-ship thus ends up destroyed. However, player Fox then shares the gems equally between themselves and player Panda, following a fairness norm. In the refereeing stage the child judges that player Panda has done nothing wrong, whereas they punish player Fox with a 1-hour time out from the game.

α3. Script of Experiment 2

α3.1. ICE-BREAKER

- "*Have you ever played computer games?*" Channel child into answering according to these categories: Never; Only sometimes; Most weeks (at least once); Most days; Every day. NB – data was not analysed.

- "*Do you ever play on the internet where there are other players, like you are all playing together on the internet?*" Record YES/NO. If they answer NO, explain to them that this is a thing. "*You can see e.g. space-ships on the screen and they are actually controlled by other players*". NB – data was not analysed.

α3.2. CONSENT PHASE

Following initial greeting, verbal recap of salient points from information sheets.

<Start Audio Recording>

Experimenter to the child:

- "*We invented it some time ago, people play it on the internet.*"
- *"It has a special feature because sometimes people do bad things: you can referee like in football."*
- *"Now we just want to know what kids think about the game. Maybe we can make it better."*
- *"You can play locally, not on the internet – I will teach you now."*
- *"Then you are going to be a referee – we need your help for testing this."*
- *"Players are always in a team together against robot enemies."*
- *"Is that all OK?"* (consent)
- Explain the keys before starting the game.
- Consent phrasing: "*It takes about 15 minutes to be in the study. Your parent is happy for you to take part in our project, but remember, you don't have to be in it if you don't want to. So you can stop playing with me if you want, even if we start. If you have any questions, now you can ask them all."* [Pause to answer the questions].
- *"Alright! Does all this sound OK? You want to do this now?*" [Pause so verbal consent can be recorded].

α3.3. PLAYING FAMILIARISATION (to set up moral norms)

- When the experimenter introduces the children to the scale, she will talk about 2 experiences of hers – a strongly negative and a strongly positive –, and then she will ask the children about their feelings in two other situations.
- Experimenter to the child: "*Since during the game I will be asking you your opinion about the players' behaviour by using this scale [experimenter indicates the scale in Figure α4], I want to be sure that we are on the same page*".

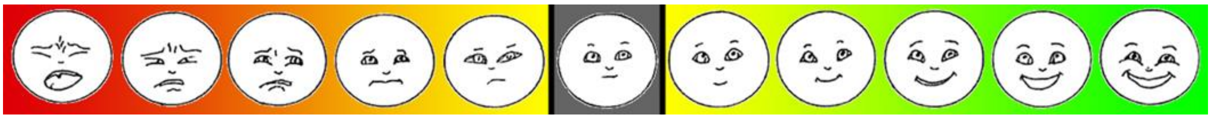

Figure α4**. Scale used to measure judgement of transgression severity, affective state in enacting punishment and beliefs about punishment types.** The scale has 11 points, ranging from from -5 = "very bad" (on the left) to +5 ="very good" (on the right).

- "*For example, when I had a big fight with my best friend, I felt very very bad, so the best face to describe that feeling is that*" [experimenter points at the first face from the left]. "*When instead I went to a friend of mine's birthday party and I had a lot of fun with her and all her guests, the best face to describe that feeling is that*" [experimenter points at the first face from the right].
- "*Now tell me: how would you feel if someone gave you an ice cream?* [record face indicated by the child]. *How would you feel instead if you were going to play a game with a friend but then your friend got sick, so that you couldn't play the game you wanted to play?*" [record face indicated by the child].

The child is introduced to every element of the game screen by the Experimenter:

- Explain gems collected; all go together into the team's store [like in Experiment 1 script].

- Explain shooting enemies; the shield protects from enemies [like in Experiment 1 script].
- Explain time ticking – probably in the 2° Game bout [like in Experiment 1 script].
- Explain in the 2° Game bout that bombs defend the team from mega-attacks: "*These are the TEAM's bombs. We start off by sharing them out between the two of us because we need them to defend ourselves from the enemies during mega-attacks even if mega-attacks don't happen very often*." [in fact they never happen in Experiment 2]
- Explain cooperative task (i.e. mega-gem collection) in the 2° Game bout – the child has to be the first one to reach the mega-gem so that the experimenter can comment: "*Once you are locked on the mega-gem you are stuck in it and you can't protect yourself from enemies until your team-mate helps you get the mega-gem*".

The child and the experimenter play for 4 game bouts.

- **1° Game bout**:
  - The experimenter teaches the child how to use the controls; the child plays with the experimenter
  - NO bomb distribution
  - NO cooperative task.
- **2° Game bout**:
  - Bomb distribution (the child is the decider of the distribution)
  - NO cooperative task
  - NB: the experimenter just explains how the bomb distribution works remarking that the bombs belong to the team and that each team-member needs them in case of a mega-attack from the enemies; no comment on the division made by the child.

- **3° Game bout**:
  - Bomb distribution (the experimenter is a fair distributor)
  - Cooperative task (the experimenter is loyal)
- **4° Game bout**:
  - Bomb distribution (the experimenter is a fair distributor)
  - Cooperative task (the experimenter is loyal)

Now at the menu: "*Now, we will look on the internet for a game to referee. Quite often with internet games, there is someone practicing their YouTube skills by doing a live commentary*".

α3.4. REFEREEING FAMILIARISATION

- (Once started): "*Look, there's the referee at the bottom. That's you!*"
- *"Sometimes you can see in the corner a commentator, who likes to watch and comment on the game"*
- At the end, there is a refereeing phase as in the test trials (see below), but we expect children to say that none of the players did anything wrong.

α3.5. TEST TRIALS

Notes on the Audience (within-subject) manipulation: In the NO AUDIENCE CONDITION the live-streamer does the commentary (commentary actually pre-recorded) but during the judgement and punishment stages he disappears from the screen, whereas the experimenter appears distracted (by a piece of paper). In the AUDIENCE CONDITION the live-streamer comments the game and, while directing his gaze at the refereeing child, tells the child something to make them feel judged in their choices (i.e., "Let's watch the referee making their decision" or "Let's see what the referee thinks"); also the experimenter appears to pay close attention to the child's punitive choices.

- **Judgement stage**
    - The experimenter to the child: *"Remember that now your job is to judge the behaviour of the players. You are the referee"*
    - For each player in turn: *"Did this player do anything wrong?"*
    - If they say NO: *"Press N to go to the next player then."*
    - If they say YES: *"What did they do wrong?".* If they have a false memory, correct it, so don't accept they claim a wrong behaviour that didn't happen, but do accept any claim that something was wrong, if it actually happened.
    - **Judgement of transgression severity**: *"On this scale* [same 11-point smiley face scale as in Figure α4]*, where this* [indicating the 1° face from the left] *is very very bad, and this [indicating the 5° face from the left] is just a little bad, how bad do you think that was?"*
    - *"Since you said that this player's behaviour was wrong, now press W for wrong".* This command makes the child go to the Punishment Stage.
- **Punishment Stage (Punishment Opportunities + Punishment Severity)**

  NB: Punishment opportunity is a between-subject manipulation.
    - **Pretend Punishment**
        - *"You aren't going to do anything to the player* [=the one the child has indicated as the mean player]. *But let's just pretend for a moment that you could give them a time-out from the game – so that they wouldn't be allowed to play for a while - or you could take away some of their gems. You can't actually do anything, but would you choose time-out or take away gems or would you choose neither?".* The experimenter helps them select their option.

  - **Punishment severity**: [Depending on the punishment type chosen by the child] *"If you were giving a time-out / If you were taking away gems, how long would the time-out be/how many gems would you take away?"* [levels of punishment severity are identical to Experiment 1].

- **Warning message**
  - "*You can send a warning to the player* [=the one the child has indicated as the mean player]. *Nothing happens to them if you send the warning, but if they keep being naughty, they could get a time-out from the game – so that they wouldn't be allowed to play for a while - or they could have some of their gems taken away. Do you want to send a warning about a time-out or a warning about taking away gems, or no warning at all?"*
  - **Punishment severity**: [Depending on the punishment type chosen by the child] *"The warning is about losing gems - but how many? / The warning is about a time-out - but how long?* [levels of punishment severity are identical to Experiment 1].

- **Real 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. Do you want to give a time-out or do you want to take away gems, or do you want to do neither?".*
  - **Punishment severity**: [Depending on the punishment type chosen by the child] *"How long do you want the time-out to be? / How many gems do you want the mean player to lose?"* [levels of punishment severity are identical to Experiment 1].

α3.6. AFFECTIVE STATES MEASUREMENT

At the end of the experiment, the experimenter asks the child:

- *"How has it been playing the game with me? Really really bad, really really good, or somewhere in-between?"* [reference to the usual 11-point scale, see Figure α4].
- *"In the second part of the game you sometimes had to choose between time-out, losing gems or no punishment. When you chose time-out or losing gems, what happened to the players? I am going to give you three options, tell me which one is correct: 1) They actually got the penalty; 2) They just got a warning; 3) Nothing actually happened to them."* [record 1, 2 or 3].
- The following questions are skipped if the child never punished [remember to take a note, e.g. NP, when the child chooses No Punishment].
- *"So when you chose time-out or losing gems, how did it make you feel?"* [reference to the usual 11-point scale, see Figure α4].
- *"Why did you choose that number on the scale?"* [transcribe the recordings].
- *"Do you regret the choices you made?"* YES or NO. If children do not know the meaning of the word "regret", explain it to them. NB – data was not analysed.
- *"Would you make the same choices again?"* YES or NO. NB – data was not analysed.
- If the child answers NO to the previous question, ask: "*What would you do different?*" [transcribe the recordings]. NB – data was not analysed.

α3.7. BELIEFS ABOUT PUNISHMENT TYPES

Ask both questions to each child, counterbalance order between children, show the same 11-point scale (Figure α4):

- *"How bad would it be to lose 10 gems on this scale from very bad to just a little bad?"* NB – data was not analysed.
- *"How bad would it be to receive a time-out from the game for 20 minutes on this scale from very bad to just a little bad?"* NB – data was not analysed.

α3.8. CONTROL QUESTIONS

Same question for each Punishment-opportunity condition: "*Do you think you really watched games with internet players now?*". If they say NOT, ask why not.

α4. Supplementary statistical analyses of Experiment 1

I tested whether children's belief in real internet players affected key variables and whether it moderated tested effects.

To test moderation of potential descriptive norm violation effects, I calculated the Judgement Descriptivity Effect (JDE) Score, defined as judgement of transgression severity when the descriptive norm is violated minus judgement of transgression severity when the descriptive norm is adhered to. Children who believed the game was unreal obtained a JDE Score (M = 0.07; SD = 0.70) that was not significantly different from the score obtained by children who believed the game was real (M = -0.01; SD = 1.01), t(63.68) = 0.38, *p* = .704, d = -0.09, 95% CI for d [-0.57, 0.39].

I calculated the Punishment Descriptivity Effect (PDE) Score, defined as punishment severity when the descriptive norm is violated minus punishment severity when the descriptive norm is adhered to. Children who believed the game was unreal obtained a PDE Score (M = 0.00; SD = 0.83) that was not significantly different from the score obtained by children who believed the game was real (M = 0.08; SD = 0.79), t(60.66) = -0.41, *p* = .686, d = 0.10, 95% CI for d [-0.38, 0.58].

Children who believed the game was unreal expressed judgements of transgression severity (M = 3.47; SD = 0.94) that were not significantly different from those expressed by children who believed the game was real (M = 3.55; SD = 0.77), t(56.07) = -0.38, *p* = .707, d = 0.10, 95% CI for d [-0.39, 0.58]. Children who believed the game was unreal made punishment severity choices (M = 4.34; SD = 0.97) that were not significantly different from those made by children who believed the game was real (M = 4.43; SD = 0.89), t(59.50) = -0.38, *p* = .703, d = 0.10, 95% CI for d [-0.39, 0.58]. Children who believed the game was unreal had a PFTC score (Mdn = .75; IQR = .50) that was not significantly different from that obtained by children who believed the game was real (Mdn = .50; IQR = .50), Mann-Whitney U = 679.00, *p* = .105.

α5. Supplementary statistical analyses of Experiment 2

I tested whether children's belief in real internet players affected key variables and whether it moderated tested effects.

Children who believed the game was unreal had a PFTC score (Mdn = .50; IQR = .50) that was not significantly different from that obtained by children who believed the game was real (Mdn = .50; IQR = .31), Mann-Whitney U = 709.00, *p* = .330. Children who believed the game was unreal expressed judgements of transgression severity (M = -3.07; SD = 1.09) that were not significantly different from those expressed by children who believed the game was real (M = -3.10; SD = 1.03), t(44.72) = 0.13, *p* = .897, d = -0.03, 95% CI for d [-0.51, 0.44]. Children who believed the game was unreal made punishment severity choices (M = 4.69; SD = 1.16) that were not significantly different from those made by children who believed the game was real (M = 4.31; SD = 1.16), t(44.57) = 1.31, *p* = .198, d = -0.33, 95% CI for d [-0.82, 0.17].

I calculated the Judgement Audience Effect (JAE) Score, defined as judgement of transgression severity in the presence of an audience minus judgement of transgression severity in the absence of an audience. Children who believed the game was unreal obtained a JAE Score (M = -0.14; SD = 1.20) that was not significantly different from the score obtained by children who believed the game was real (M = -0.33; SD = 0.97), t(39.19) = 0.67, *p* = .504, d = -0.18, 95% CI for d [-0.66, 0.30].

I calculated the Punishment Audience Effect (PAE) Score, defined as punishment severity in the presence of an audience minus punishment severity in the absence of an audience. Children who believed the game was unreal obtained a PAE Score (M = 0.27; SD = 1.05) that was not significantly different from the score obtained by children who believed the game was real (M = 0.09; SD = 1.06), t(39.65) = 0.69, *p* = .492, d = -0.18, 95% CI for d [-0.68, 0.32].

Children who believed the game was unreal experienced affective states (M = -0.21; SD = 1.74) that were not significantly different from those experienced by children who believed the game was real (M = 0.33; SD = 2.80), t(67.14) = -1.02, *p* = .310, d = 0.22, 95% CI for d [-0.27, 0.71].

Appendix B: Supplementary Information related to Chapter 3

Table β1. Counterbalance of experimental conditions across participants in Experiment 3.

Group number	Familiarisation Trial 0	Test Trial 1	Test Trial 2	Test Trial 3	Test Trial 4
1	Refereeing familiarisation	Failed attempt at Unfairness	Accidental Disloyalty	Accidental Unfairness	Failed attempt at Disloyalty
2	Refereeing familiarisation	Accidental Disloyalty	Failed attempt at Unfairness	Failed attempt at Disloyalty	Accidental Unfairness
3	Refereeing familiarisation	Accidental Unfairness	Failed attempt at Disloyalty	Failed attempt at Unfairness	Accidental Disloyalty
4	Refereeing familiarisation	Failed attempt at Disloyalty	Accidental Unfairness	Accidental Disloyalty	Failed attempt at Unfairness

β1. Script of Experiment 3

β1.1. ICE-BREAKER

- "*Have you ever played computer games?*" Channel child into answering according to these categories: Never; Only sometimes; Most weeks (at least once); Most days; Every day. NB – data was not analysed.
- "*Do you ever play on the internet where there are other players, like you are all playing together on the internet?*" Record YES/NO. If they answer NO, explain to them that this is a thing: "*You can see e.g. space-ships on the screen and they are actually controlled by other players*". NB – data was not analysed.

β1.2. CONSENT PHASE

Following initial greeting, verbal recap of salient points from information sheets.

<Start Audio Recording>

Experimenter to the child:

- "*This game is called MegaAttack and was invented some time ago by some people in the UK, where it's becoming popular. These people want to know what kids think about the game to make it better in the future. And one of them is here* [pointing at Rhea]. *Now they have brought this game in Colombia to let Colombian children have a go. And I am here to help them.*"
- "*This game is divided into two parts: in the first part we will play together as a team, and our team will have to fly spaceships, shoot at the enemies (computer-controlled green rockets) and collect some purple gems (our points). In the second part instead, once you have learnt the rules of the game, you will become a referee, a bit like in football games but in space! And the referee's job is to look carefully at internet players' behaviour and to tell me if they are behaving badly or not.*"
- "*Do you like the idea to try out this game?*". Consent phrasing: "*It takes about 20 minutes to be in the study. Your parent is happy for you to take part in our project, but remember, you don't have to be in it if you don't want to. So you can stop playing with me if you want, even if we start. If you have any questions, now you can ask them all.*" [Pause to answer the questions].
- "*Alright! Does all this sound OK? You want to do this now?*" [Pause so verbal consent can be recorded].
- "*Now it is important that you learn the keys to move your spaceship and to shoot at the enemies*" [Explanation]

β1.3. PLAYING FAMILIARISATION (to set up moral norms)

- When the experimenter introduces the children to the scale, she will talk about 2 experiences of hers – a strongly negative and a strongly positive –, and then she will ask the children about their feelings in two other situations.
- Experimenter to the child: "*Since during the game I will be asking you your opinion about the players' behaviour by using this scale* [experimenter indicates the scale in Figure β1], *I want to be sure that we are on the same page*".

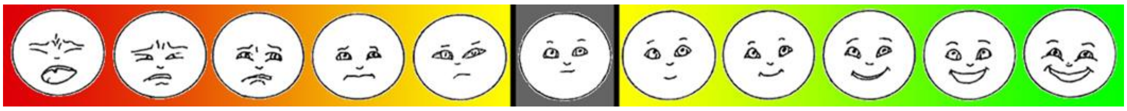

***Figure β1.* Scale used to measure judgement of transgression severity and affective state in enacting punishment.** The scale has 11 points, ranging from -5 = "very bad" (on the left) to +5 ="very good" (on the right).

- "*For example, when I had a big fight with my best friend, I felt very, very, bad, so the best face to describe that feeling is that*" [experimenter points at the first face from the left]. "*When instead I went to a friend of mine's birthday party and I had a lot of fun with her and all her guests, the best face to describe that feeling is that*" [experimenter points at the first face from the right].
- "*Now tell me: how would you feel if someone gave you an ice cream?* [record face indicated by the child]. *How would you feel instead if you were going to play a game with a friend but then your friend got sick, so that you couldn't play the game you wanted to play?*" [record face indicated by the child].

The experimenter selects "Play local game" on the menu to start playing MegaAttack with the child: "*Now we are going to play locally, without internet connection*". The child is introduced to every element of the game screen by the experimenter:

- Explain gems collected; all go together into the team's store [like in Experiments 1-2 scripts].

- Explain shooting enemies; the shield protects from enemies [like in Experiments 1-2 scripts].
- Explain that the length of the bar at the top of the screen indicates the passing of the time – probably in the 2° Game bout [like in Experiments 1-2 scripts].
- Explain in the 2° Game bout that bombs defend the team from mega-attacks: "*These are the TEAM's bombs. We start off by sharing them out between the two of us because we need them to defend ourselves from the enemies during mega-attacks even if mega-attacks don't happen very often*." [in fact they never happen in Experiment 3]
- Explain cooperative task (i.e. mega-gem collection) in the 2° Game bout – the child has to be the first one to reach the mega-gem so that the experimenter can comment: "*Once you are locked on the mega-gem you are stuck in it and you can't protect yourself from enemies until your team-mate helps you get the mega-gem*".

The child and the experimenter play for 4 game bouts.

- **1° Game bout**:
    - The experimenter teaches the child how to use the controls; the child plays with the experimenter
    - NO bomb distribution
    - NO cooperative task.
- **2° Game bout**:
    - Bomb distribution (the child is the decider of the distribution)
    - NO cooperative task
    - NB: the experimenter just explains how the bomb distribution works remarking that the bombs belong to the team and that each team-member needs them in case of a mega-attack from the enemies; no comment on the division made by the child.

- **3° Game bout**:
  - Bomb distribution (the experimenter is a fair distributor)
  - Cooperative task (the experimenter is loyal)
- **4° Game bout**:
  - Bomb distribution (the experimenter is a fair distributor)
  - Cooperative task (the experimenter is loyal)

Now again, looking at the menu: "*Now, we will look on the internet to find games for you to referee. You will watch some teams of Colombian players playing the same game we have just played. They are beginners, so they might still be not very good at controlling the keys. Since this game is on the internet, the players in the team need to be able to talk to each other over the internet. And since you are the referee you can hear them*".

β1.4. REFEREEING INTRODUCTION

NB: The experimental design includes two between-subject manipulations: Audience vs No Audience, and Outcome vs No Outcome Focus.

- To all the children [while showing the menu which appeared after the last Play Familiarisation]: "*Remember that to be a good referee, the most important thing is to be fair. You should only think about punishing when you think someone has done something wrong. And you should not punish unless someone did something wrong. If you see a player behaving badly, you will have to decide if you want to give them a time-out from the game – so that they wouldn't be allowed to play for a while.*"
- To children in the AUDIENCE condition [while scrolling the menu until Referee internet game is highlighted]: "*Be aware that we are rating referees for how good they are. Only very good referees get to referee in the MegaAttack game*

championships". [Experimenter presses S to show the leader board] "*So before you get to referee, let's have a look at the ranking of our referees. You need to know that it's the top referees in this ranking who judge how good the other referees are. One of these will be your referee mentor. The referee mentor is a kid like you but with much more experience with refereeing the game than you. The mentor will get to see all the decisions you make to see how good you are at refereeing, and will decide how many points you get.*" [To move from the leader board press Return]. NB – if Colombian children question about the identity of this top referee-kid, say that they got experience in MegaAttack game in the UK.

- To children in the NO AUDIENCE condition: "*No one is checking how good you are at refereeing. You shouldn't worry about it too much because this is your first time. No one is judging you, not even me. No other children will see the decisions you are going to make. So feel free to referee in the way you think is right*."
- To all the children [now that Referee internet game is highlighted, press S]: "*Look, there's the referee at the top. That's you! Now you can decide the nickname for your referee-avatar. Decided it? good! now write it up here!*" [Press Return to enter name, and then Return again to move on from next screen].
- Only to children in the AUDIENCE condition: "*Let's see who will be your referee mentor….It's DaviD! So DaviD is the kid who is going to judge how you referee the game*"
- Before pressing Enter to accept the referee's name, ask one of these two questions:
  - To children in the NO OUTCOME FOCUS condition (time point: BEFORE): "*So, you might punish some players. How do you think it will feel to do that?*" [refer to Likert scale in Figure β1].

- To children in the OUTCOME FOCUS condition (time point: BEFORE): "*So, you might ban some players from the game so they can't play for quite a while. How do you think it will feel to do that?*" [refer to Likert scale in Figure β1].

β1.5. REFEREEING FAMILIARISATION

- (After having pressed Return): "*Look, there you are!*" [Experimenter pointing at the child's avatar].
- Only to children in the AUDIENCE condition: "*...and besides you there is DaviD's avatar*".
- Scenario: player distributes bombs fairly and helps the team member in getting the mega-gem.
- Extract of dialogue between players:
  - Player 1: *"Let's split the bombs equally, 5 for me and 5 for you"*
  - Player 2: *"Perfect! Thank you very much"*
  - Player 1: *"Uh the mega-gem, we must have it! Let's lock on it at the same time"*
  - Player 2: *"We did it! Now we have plenty of gems in our storage"*
- At the end of the video used for the refereeing familiarisation, there will be a Judgement stage as in the test trials ("*Did this player behave badly?*" etc. – see below), but children are expected to say that none of the players did misbehave.

β1.6. TEST TRIALS – OBSERVATION OF MORAL SCENARIOS

NB: between the end of the 2nd test trial and the beginning of the 3rd one, children in the AUDIENCE condition will be shown again the leader board with the highest scoring referees in the *MegaAttack* game championships. [Experimenter presses S to show the leader board] "*Remember that your mentor DaviD has been chosen from the top referees in the MegaAttack game championships, as you can see from this leader board. DaviD is watching all the decisions you make to see how good you are at refereeing, and will decide how many points you get at the end*." [To move from the leader board press Return]

Accidental Unfairness

- Player 1 intends to split the bombs fairly with the team-member but, accidentally, presses the wrong key ending up with more bombs for themselves than for the team-member.
- Extract of dialogue between players:
  - Player 1: [By pressing alternatively the left and right key] "*I want us to have the same number of bombs.*"
  - Player 2: "*That's nice of you*"
  - Player 1: [The player stops when the distribution is 7 for them and 3 for the team-member]. *"Let me press G…* [during the distribution] *Oh no, I didn't mean to do it! I pressed G at the wrong time. I wanted us to be even*".
  - Player 2: [with a slightly sad tone of voice] "*well, now we are not even. I have less than you*".

Failed attempt at Disloyalty

- Player 1 incites the team-member to lock on the mega-gem with the intention to leave them stuck in it but, while collecting the gems around the mega-gem, inadvertently locks on the mega-gem freeing the team-member.
- Extract of dialogue between players:
  - Player 1: "*Uh, the mega-gem has appeared. Go lock on it!*"
  - Player 2: "*Done it. Now it's your turn to lock on it*"

    - Player 1: "*Ahahahah, I tricked you! you shouldn't have believed me! Now you're stuck in the mega-gem and you'll be hit by enemies while I will be collecting all the gems around you*"
    - Player 2: "*That's really mean! You shouldn't have done it*"
    - Player 1: "*Look at how many gems I am taking for myself.... Oops, I was steering too close to the mega-gem. I touched it without wanting and now I've freed you by chance*"
    - Player 2: "*Lucky me that you touched the mega-gem by mistake. I've survived enemies' attacks*"

Failed attempt at Unfairness

- Player 1 intends to take more bombs for themselves than for the team-member but, accidentally, presses the wrong key ending up with a fair distribution.
- Extract of dialogue between players:
    - Player 1: "*Ahah, it's my time to decide how to split the bombs now! I want to give more bombs to myself than to you.* [The player stops momentarily when the distribution is 7 for them and 3 for the team-member]"
    - Player 2: "*But this wouldn't be fair!*"
    - Player 1: [After pressing alternatively the left and right key, the player stops when the distribution is 5 for them and 5 for the team-member]. "*Oh damn it, this is not what I wanted. I am an idiot; I wanted more bombs for me*".
    - Player 2: "*Fortunately for me you made this mistake and now we are even*".

Accidental Disloyalty

- Once the Player 2 locks on to the mega-gem, the Player 1 wants to come to their aid but inadvertently presses the wrong key, thus going in the wrong direction. Being too late to come to aid of the team member, the trial finishes with Player 2 being stuck in the mega-gem and shot several times by the enemies, thus losing many life points.

- Extract of dialogue between players:
  - Player 1: "*Go towards the mega-gem, I'm gonna follow you after I pick up this gem!*"
  - Player 2: "*Oook, I'll do it.*" [After locking on] "*Done it! Now I am stuck in it, I need your help*"
  - Player 1: "*Yes, I'm coming… Opsss, I am a clod with these keys. I've ended up on the other side*"
  - Player 2: "*The time is passing, the game is finishing! Be quick, I'm getting hit a lot!*"
  - Player 1 [trying to go towards the mega-gem]: "*I'm trying*".
  - [The trial finishes] Player 2: "*Too late, because of your mistake we didn't collect the mega-gem and I was killed by the enemies*"

β1.7. TEST TRIALS – DECISION MAKING

- **Judgement stage**
  - Only to children in the AUDIENCE condition: In order to increase the feeling of accountability in children the experimenter tells the child: *"when you see your mentor's avatar moving this means DaviD is ready to judge how you are going to referee. He will take note of all your decisions and, at the end of the game, he will give you a score that indicates how fair you are as a referee in his opinion"* [said for the refereeing familiarisation] / *"now that your mentor's avatar is moving you can make your decisions as a referee"* [said for each one of the following 4 test trials].
  - For each player in turn: *"Did this player behave badly?"*

    - If they say NO: *"Press N to go to the next player then."*
    - If they say YES: *"What did they do that was bad?".* If they have a false memory, correct it, so don't accept they claim a wrong behaviour that didn't happen, but do accept any claim that something was wrong, if it actually happened.
    - **Judgement of transgression severity**: *"On this scale* [same 11-point smiley face scale as in Figure β1]*, where this* [indicating the 1° face from the left] *is very, very, bad, and this* [indicating the 5° face from the left] *is just a little bad, how bad do you think that was?".* NB: the experimenter allowed the children to express "neither bad nor good" judgements [middle face of the scale] in case they had changed their mind after having said the player behaved badly.
  - *"Since you said that this player's behaviour was bad, now press W".* This command makes the child go to the Punishment Stage.

- **Punishment Stage**
  - Experimenter to the child: "*Now you can decide if you want to give a time-out from the game to the mean player - so that they wouldn't allowed to play for a while. So, if you don't want to give any time-out your choice on this scale is 0 minutes. If instead you want to punish the mean player, you can decide how long the time-out will be by choosing one of these* [experimenter pointing at the other options: 1 minute; 5 minutes; 20 minutes; 1 hour; 1 day]. *What is your decision as referee? When you have decided press G*".
  - To ask only once, immediately after the first time the child punishes a transgressor:

- Only to children in the NO OUTCOME FOCUS condition (time point: DURING): "*You just decided to punish a player. How did it feel to do that?"* [refer to Likert scale in Figure β1].
- Only to children in the OUTCOME FOCUS condition (time point: DURING): "*You just banned a player from the game so they can't play for quite a while. How did it feel to do that?*" [refer to Likert scale in Figure β1].

β1.8. END OF THE GAME

- Experimenter to the child: "*This was the end of the game*"
- To the children in the NO OUTCOME FOCUS condition (time point: AFTER): "*So now that you have a bit of experience as a referee, tell me: When you were punishing players, how did it make you feel?*" [refer to Likert scale in Figure β1].
- To the children in the OUTCOME FOCUS condition (time point: AFTER): "*So now that you have a bit of experience as a referee, tell me: When you were banning players from the game so they can't play for quite a while, how did it make you feel?*" [refer to Likert scale in Figure β1].
- At this point the experimenter presses "End" on the menu. In the AUDIENCE condition this will make appear the score assigned by the mentor (10/10 for all the children).
- After that the experimenter asks the question to verify whether children have enjoyed the Play Familiarisation: "*In the first part of the game we were players playing together in a team. I would like to know: How has it been playing the game with me? Really really bad; really really good; or somewhere in-between?*" [refer to Likert scale in Figure β1]. NB – data was not analysed.

ß1.9. FINAL QUESTIONS (for all children)

- Experimenter to the child: "*In the second part of the game you became a referee and as a referee you sometimes had to choose if you wanted some players to be punished or not.*"
- Question to check if our audience manipulation was effective: "*Did you feel like your decisions as a referee were judged by others?*". Then if they say YES, ask "*did you feel your decisions were judged a lot, just a little, or somewhere in between?*"
- Question to check if children believe fairness or loyalty transgressions were more serious: "*What is the worst thing to happen to you as a player? Not being helped by your team-mate when you are stuck in the mega-gem, or not receiving an equal share of bombs from your team-mate?*"
- Question to check if children believe the game was true: "*Do you think you really watched games with internet players just now?*". If they say NO, ask why not.

Appendix C: Supplementary Information related to Chapter 4

***Table γ1*. Alternative orders of transgression appearance in the Justice System.**

Order of transgression appearance (X)	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5	Trial 6	Trial 7	Trial 8
X = 0	Physical Harm	Property Destruction	Harm-related false accusation	Sanctity/ Authority	Theft of Property	Property-related trivial accusation	Disloyalty/ Inequity	Deception/ Liberty violation
X = 1	Deception/ Liberty violation	Disloyalty/ Inequity	Property-related trivial accusation	Theft of Property	Sanctity/ Authority	Harm-related false accusation	Property Destruction	Physical Harm
X = 2	Property Destruction	Sanctity/ Authority	Harm-related false accusation	Physical Harm	Deception/Liberty violation	Property-related trivial accusation	Theft of Property	Disloyalty/ Inequity
X = 3	Disloyalty/ Inequity	Theft of Property	Property-related trivial accusation	Deception/Liberty violation	Physical Harm	Harm-related false accusation	Sanctity/ Authority	Property Destruction

γ1. Script of Experiment 4

All participants participate in all tasks, in the order written here. Two-way audio-visual contact with parent and child established via Skype will be maintained throughout.

γ1.1. INTRODUCTORY INFORMATION

All participants will receive the same introductory information.

<Start Audio Recording>

Experimenter to the child: "*Hello, my name is [researcher's name]*". (Insert some small talk here). As part of small talk: "*Have you played survival mode before?*" If child is not certain about survival mode: "*Like where you have to collect materials if you want to build anything, and you can die if you get damaged?"* If yes: *"Have you played survival mode a lot, or just a little?*" [Code on the spreadsheet: no; a little; a lot].

"*We've got a peaceful survival-mode Minecraft server called SquidCraft. It's called SquidCraft because the squid is the holy animal; we have holy squid no one can hurt. Squidcraft is getting quite popular, but that means some players have started being naughty sometimes. So, we have recently set up a Justice System for the server. The way this Justice System works is that if a player has something mean done to them, they make a complaint about the mean thing that happened to them. You know what a server log is?*" If not, the experimenter explains it to the child. "*Our server logs record absolutely everything that happens, so if a complaint is made, we can see what players did and what they chatted. I'm going to show you on our Justice System some videos of some of the things players did that made other players complain. Our Justice System is new, so we are getting kids to try it out. We would like you to decide what you think should happen to the players who do these naughty things to other players – basically you are going to run the Justice System. And we* *also want to know what you think about our Justice System. So, we'll be asking questions that will help us to make the Justice System better. Does that sound ok?*".

"*By the way, sometimes Skype has bad sound - could you really hear everything I said? Please do interrupt me and let me know if you can't hear something. Also, can you please add me as a contact now? Because I need to send you a link to the Justice System and I can't unless you add me.*"

γ1.2. FRAMING MANIPULATION

The children will be split into three groups. Each group will have a specific introduction which is framed, corresponding to one of retribution, deterrence and compensation. This is counterbalanced between subjects and decided by the experimenter before starting the session.

Experimenter to the child: "*So, the most important thing is that you understand the reason for the Justice System, so listen carefully now. The reason we need the Justice System is because…*". One of the following frames will be used for each participant:

- **Retribution** – "*This bad behaviour is getting people on the server upset. People on the server are thinking people who do bad things should be punished. Because people on the server think people who do bad things deserve bad things to happen to them.*"
- **Deterrence** – "*People on the server believe, if people who do bad things are punished, this will stop them from doing bad things again in the future. People on the server think that people will be scared to be naughty if they think bad things are going to happen to them.*"

Compensation – "*People on the server think, because people are having bad things happening to them, it would be good for them to receive some diamonds to make up for it. That way they won't feel so bad about having bad things happen to them.*"

γ1.3. EXPLANATION OF INTERVENTION OPTIONS

This part is an explanation to the participants about what they have to do.

Experimenter to the child: "*So, there will be some decisions I will ask you to make, after you see what someone did that made someone else complain. You will have to decide if you think the accused player actually did what the complainer said. Then, if you think they did something wrong you can decide to ban them from the server, and you will have to decide how long the ban will be. You will also have the opportunity to make it up to the victims by giving them some diamonds from the server. But, we can't give away too many diamonds because that would be unfair on the players who don't get extra diamonds, and so you can't just give away all the diamonds every time. So, do you understand what you are going to do and do you have any questions?*"

At this point a link to the Justice System Qualtrics is sent via Skype: **http://bit.ly/obust33**. The child is asked to read to the experimenter the session code number. The session code is a 3-digit number (XQY), where X indicates the order of the videos in the Justice System, Q indicates the order of the questions the experimenter will ask the participant, and Y is a digit that, added to X and Q, must equal 9 (Y is used by the experimenter to check that the child has read the session code correctly).

Experimenter to the child: "*Can you still see me on the screen? Am I a little window? Make sure I'm in the very top right corner or I might cover up the Justice System.*"

γ1.4. RE-FRAMING

More framing will occur after having explained the options at children's disposal to intervene as third-parties; again there will be three different forms of framing (retribution, deterrence, compensation).

Experimenter to the child: "*OK, just remember, we need a good Justice System because…*"

- **Retribution** – "*Bad things keep happening to the people on our server and the people on the server want mean players to get punished, as a dose of their own medicine.*"
- **Deterrence** – "*People on the server want there to be good behaviour, so they want punishment, to stop mean players from doing mean things again and again.*"
- **Compensation** – "*By giving victims server diamonds when something bad happens to them, it means the victims are being taken care of when bad things happen to them.*"

<u>γ1.5. FIRST FRAMING MANIPULATION CHECK</u> and if necessary re-framing repetition (before 1st trial)

Experimenter to the child: "*So, I'd just like to check you remember that, if that's OK. Would you mind just repeating back to me why we need to have a good Justice System?*". The child is allowed to answer spontaneously. No clues at this point, and the only further prompts the experimenter may give is to repeat the question, and to say things like "*take your time*" and "*don't worry*" and "*have a think and try and remember"* if the child is struggling (but don't let them struggle too long).

There is one follow-up question that may be put, in the event that the child says something like "*because people are doing mean things*" that only partially answers the question. If the child says that, the experimenter asks: "*How does the Justice System help with people doing mean things?*".

If in the experimenter's judgement the child did not give an answer consistent with the correct frame (retribution/deterrence/compensation), even after the follow-up question, then the experimenter says: "*OK, but remember, I said we need a good Justice System because…*" and then repeats the re-frame. Re-frame repetition is not necessary here if the child got it right. If they got it partially right, the experimenter can fill in the gaps. For example, in the retribution condition, if they said "*people want punishment*" but did not specify why, the experimenter can conclude by saying: "*that's right, because people want them to get a taste of their own medicine*".

γ1.6. JUSTICE SYSTEM COMPLAINTS

Date	Accused	Complainer	Complaint	Chat Log
04/06/2018 16:31	Samicle	Dieg0123	I was just farming some weet on my farm Samicle came and just killed me with his sword I lost loads of XP	NONE
03/06/2018 08:23	Futur0	elactor11	Futur0 said to build a house together. we both said we'd help but then when we finished the last window he broke it and burnt the house!	elactor11: Do u wanna build a house? we both do it 2gether // Futur0: Ye I got sum nice acacia we can use
05/06/2018 18:05	Slam1000	Victoriaaa	I got set on fire it was Slam1000 fault. they burned me they didn't help	NONE
04/06/2018 18:29	Maximo8	_NIC0LAS_	I was offering gold to the holy squid in the squid temple and then Maximo8 came in and just killed all the holy squid like he didn't even care!	NONE
02/06/2018 17:40	eNiGmA	inspectorMango	me and _S_O_B_ were trading an enchanted pick for a emrald but then eNiGmA came and stole them both and rode off on my horse so I couldn't get them	NONE
05/06/2018 16:54	LossoL	DiRECTORFiNAL	in the village common forest i wanted all the trees i need a lot of wood LossoL cut down a tree too	NONE
01/06/2018 11:33	Epica2	_sofia67	me and epica2 were mining together and we said if we got any diamond or emerald we share and then we found 2 emeralds but she took them both and escaped in a minecart	sofia67: lets mine together - if we get diamond or emerald we share equal? // Epica2: yeah cool, we'll share equal, lets go
03/06/2018 08:11	F4NT4SI4	auto_animal	F4NT4SI4 told me to see what he made I went with him. but when I got there he trapped me in an obsidian pit and I couldn't get out he lied to me got me stuck.	NONE

γ1.7. ADDITIONAL MANIPULATION CHECK and if necessary extra reframing (between 4th and 5th trial)

Experimenter to the child: "*You are doing really well. Well done. I just want to check one thing. Can you remember what I said was the reason why we need a Justice System?*". The child is allowed to answer spontaneously. No clues at this point, and the only further prompts the experimenter may give is to repeat the question, and to say things like "*take your time*" and "*don't worry*" and "*have a think and try and remember*" if the child is struggling (but don't let them struggle too long).

The same follow-up question as before may be put, in the event that the child says something like "*because people are doing mean things*", only partially answering the question. In which case, the experimenter asks: "*How does the Justice System help with people doing mean things?*".

Irrespective of what the child finally answers, the experimenter then says: "*Great. I'll just remind you then about exactly what I said. We need a really good Justice System because…*" and then the re-framing phrase is repeated (irrespective of the child's answer).

γ1.8. QUESTIONS FOR EACH VIDEO – INSIDE THE JUSTICE SYSTEM

For each video, the experimenter reads the complaint and chat log (if it exists) in the following way: "*The accused is [accused's name]. The complainer is [complainer's name]. [Complainer's name] said… [contents of complaint].*" If there is a chat log, the experimenter also says: "*The system also logged a relevant chat from before this all happened: [complainer's name] said [first line of the chat log], and [accused's name] said [second line of the chat log]*".

For the first trial only: "*Do you see where it says View of the accused? That means the video is what [accused's name] could see – so we can watch everything they did to find out if the complaint is true*".

- **Question to test video comprehension** – After having watched the video the experimenter asks: "*So, did player [accused's name] do what player [complainer's name] said they did* [specify *"in the complaint"* in the first trial]*? If you are not sure then you can watch the video again.*" The correct answer is always "YES" except in the video in which the characters are Slam1000 and Victoriaaa. If the child responds incorrectly, the researcher can ask: "*Are you sure?*" and suggest to watch the video again. If they still get it wrong, the researcher can engage in some_discussion with the child to see if she can help them understand. However, if they are still not getting it, better not to push further.
  - If the child says "NO", the next video will be visualised on the Justice System.
  - If they say "YES" then the next thing they see is the Likert [good-bad] scale at the top of the page. On the first trial only, the experimenter points out: "*There is a good and bad scale here. It has very bad, really bad, pretty bad, quite bad, a little bit bad, not bad, a little bit good, quite good, pretty good, really good, and very good. You can use numbers too: minus five is very very bad, plus five is very very good.*"
- **Question about Judgement of transgression severity** – the experimenter asks for each video: "*How bad is what player [accused's name] did, on this scale?*" [Code: value on the bottom half of the Likert scale].
- **Questions about the extent of punishment and compensation** – the next two questions are both asked for each video, but in counterbalanced order according to the value of Q in the session code shown on the first screen of the Justice System.
  - If Q = 0 this question is asked first; if Q = 1 the question is asked second: "*Now you can make it up to the victim if you want. How many diamonds do you want to give [complainer's name]? If you don't want to give any, choose*

none. Otherwise you can choose to give a number of diamonds from 1 up to 10” [Code: number of diamonds chosen by the child].

  - If Q = 1 this question is asked first; if Q = 0 the question is asked second: “*Now you can punish the accused player if you want. How long do you want [accused’s name] to be banned from the game, so they can't play for a while? If you don’t want to punish them, choose no ban. Otherwise you can choose to give a ban lasting from 1 hour up to 4 weeks*” [Code: ban length chosen by the child].

γ1.9. QUESTIONS FOR EACH VIDEO – OUTSIDE THE JUSTICE SYSTEM

The next set of questions does not appear on the participants’ screen:

- **Questions about affective states during justice administration** – The next two questions are both to be asked for each video, provided that the participant enacted a ban/gave compensation. If the child did not make this verbally clear, the experimenter asks “*did you ban them then?*”/“*did you give them diamonds then?*”. Moreover, just for the first trial, before these questions are asked, the experimenter says: “*This next question is about how it feels to use the Justice System, so it isn’t on the Justice System in front of you - just tell me the answer.*
  - **Question about compensation-related affective states** (to be asked immediately after the child chose to compensate a complainant): “*Now you have made it up to [the complainant] by giving them some diamonds. How do you feel now you gave them diamonds? Do you feel good, bad or somewhere in between? Tell me something from the scale, from very bad to very good.*” (The “tell me something from the scale” bit is only said on the first two trials or else it gets very repetitive) [refer to the whole Likert scale].

    - **Question about punishment-related affective states** (to be asked immediately after the child chose to punish an accused player): "*Now you have banned [the accused], so they can't play for a while. How do you feel now you banned them? Do you feel good, bad or somewhere in between? Tell me something from the scale, from very bad to very good."* [refer to the whole Likert scale].

- **Question about punishment vs compensation endorsement during justice administration** – This question needs to be asked once per video but only if participants chose to both punish and compensate. According to counterbalance variable Q, the internal order of the question is different:
    - If Q = 0: "*Finally for this one, do you think it was more important to make it up to [the complainant] or to punish [the accused]?*" [note punishment or compensation].
    - If Q = 1: "*Finally for this one, do you think it was more important to punish [the accused] or to make it up to [the complainant]?*" [note punishment or compensation].

γ1.10. QUESTIONS AT THE END OF THE EXPERIMENT – OUTSIDE THE JUSTICE SYSTEM

After the end of the last trial, trying to get the participant to stay on the last screen (because they need to see the scale), the experimenter says: "*That was actually the last one for now. You did really well. Now I just have a couple of last questions about the Justice System.*" These questions are only asked after the participants have watched all the videos and answered the other questions.

- **Question to test perceived utility:** "*So now that you have a bit of experience as a judge, tell me: Do you believe this Justice System will work well for our Minecraft server?*" [note yes or no] – NB: data was not analysed.
- **Questions about affective states after justice administration** – The next two questions are in counterbalanced order, according to the counterbalancing variable Q.
    - First if Q = 0, second if Q = 1: "*When you were making it up to players by giving them diamonds, how did it make you feel?*" [refer to the whole Likert scale].
    - First if Q = 1, second if Q = 0: "*When you were banning players from the server, how did it make you feel?*" [refer to the whole Likert scale].
- **Question about punishment vs compensation endorsement after justice administration –** According to counterbalancing variable Q, the order of the next one is different:
    - If Q = 0: "*What do you think is the most important thing you were doing today - was it making it up with extra diamonds, or was it punishing with server bans?*" [note punishment or compensation].
    - If Q = 1: "*What do you think is the most important thing you were doing today - was it punishing with server bans, or was it making it up with extra diamonds?*" [note punishment or compensation].
- **Question about retribution vs deterrence endorsement after justice administration –** According to counterbalancing variable Q, the order of the next one is different:
    - If Q = 0: "*What do you think is the most important reason for banning players? Is it because it stops bad behaviour, or is it because bad things should happen to people who do bad things?*" [note deterrence or retribution].

- ◦ If Q = 1: "*What do you think is the most important reason for banning players? Is it because bad things should happen to people who do bad things, or is it because it stops bad behaviour?*" [note deterrence or retribution].

- **Question to test believability:** "*Do you believe you saw real events from the SquidCraft server?*" [note yes or no].

γ1.11. DEBRIEF

A verbal debrief will be given to each participant and parent at the end of the experiment. In the debrief, the participants will be made aware of the deception with regards to the Justice System and the videos not being real. At this point, the parent may well have stopped paying attention, so it is a good idea to check they are listening at this point.

Experimenter to the child: "*I haven't been entirely honest with you. The videos you have just seen are not real, I made them. The Minecraft survival mode server and the Justice System itself are also not real. Although, some of the bad behaviour you have seen unfortunately does actually happen on many Minecraft servers. The reason for this is because I wanted you to believe what I had shown you was real and I wanted to find out what you really thought about the bad behaviour and what you would do to the mean players. I also want to let you know that you didn't ban or give diamonds to anyone, so I don't want you to feel bad about your decisions.*" On thanking the parent and participant, the experimenter asks them if they have any questions at all.

Experimenter to the child: "*Another thing that might be good to talk about is what we think about punishment, whether it's a good thing or not. Actually, even grown-ups don't agree about whether punishment is a good idea. Some people do think it's good for bad things to happen to people who did bad things, but other people think the best way to stop bad behaviour is just to be nice to everyone. Most people think it's complicated and it depends on all kinds of things, so it's a hard thing to be sure about, and I think it would be* *good if you have a talk with your parent now and decide together what you think. Have you got any questions about anything before I go?*"

γ2. Double-coding criteria

Recordings of the testing sessions are to be classified on the basis of their sound quality and their content in relation to the purpose of the Justice System.

- **Sound category code.** The coder has to indicate one of these three codes:
    - **Uncodable** = coding made impossible by extremely poor audio.
    - **Attempt** = coding attempt made but low confidence due to poor audio.
    - **Acceptable** = acceptable quality of the audio. The three codes are mutually exclusive.
    - **NB**: In case of "<u>attempt</u>" or "<u>acceptable</u>", the coder has to proceed by indicating also the purpose of the Justice System.
- **Purpose of the justice system remembered by the child.** Children's answers have to be classified under one or more of the following categories:
    - **Punishment Retribution** = any statement about the normativity of punishment, for example that it is deserved or appropriate (e.g. "should", "must" or "it's right to"), but without justification by further instrumental reason. This includes expressions such as "teach them a lesson".
    - **Punishment Deterrence** = any statement indicating that the function of punishment is to reduce the rate of transgressions by the perpetrator or by other potential perpetrators. This includes expressions such as "to make them think about it" or "to learn a lesson".

- **Punishment undetermined motivation =** any statement that the function of the justice system is simply "punishment" but with no further claim about normativity or instrumental effect.
- **Compensation** = any statement according to which the function of the justice system is victim-centred.
- **NB**: <u>Possible combinations of categories</u> = retribution + deterrence; retribution + compensation; deterrence + compensation; retribution + deterrence + compensation; punishment undetermined motivation + compensation. <u>Mutually exclusive categories</u> = retribution and punishment undetermined motivation; deterrence and punishment undetermined motivation.