PRESCHOOLER'S EVALUATION OF AUTHORITY FIGURES' THIRD-PARTY PUNISHMENT OF A MORAL TRANSGRESSION

by

YUNRU MA

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ARTS

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Psychology)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

August 2023

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the thesis entitled:

Preschooler's Evaluation of Authority Figures' Third-party Punishment of a Moral Transgression

submitted by Yunru Ma in partial fulfilment of the requirements for

the degree of Master of Arts

in Psychology

Examining Committee:

Dr. J. Kiley Hamlin, Professor, Psychology, UBC

Supervisor

Dr. Andrew S. Baron, Professor, Psychology, UBC

Supervisory Committee Member

Dr. Susan Birch, Professor, Psychology, UBC

Supervisory Committee Member

Abstract

Punishment, despite its negative nature, plays a crucial role in fostering cooperation within human society by deterring antisocial behavior and promoting prosocial behavior in long-term social interactions. Although some evidence suggested children would consider the target's previous prosocial or antisocial actions in their socio-moral evaluation (Geraci, 2021; Hamlin et al., 2011; Lee & Warneken, 2020; Loke et al., 2011), some showed they do not (Li et al., 2020; Li & Tomasello, 2018; Van de Vondervoort, 2020). One possible explanation is that children are opposed to punishment from an ordinary citizen who is not in the position to punish. We hypothesize that children may perceive punishment as acceptable when carried out by authority figures. To explore this hypothesis, the present study investigated whether 3- and 4-year-old children would consider the context in which helping and hindering occur in their evaluations when the moral agents were depicted as holding authority status.

Contrary to our prediction, both 3- and 4-year-old children negatively evaluated the punishing police officers and positively evaluated the rewarding police officer regardless of whether the target was previously prosocial or antisocial. They preferred the rewarding police officer when asked about liking, the rightness of the action, and identifying the good police officer. We discussed possible reasons for our failure to detect children's context-dependent moral evaluations using the current study design, and proposed future direction to explore this topic. These findings contribute to our understanding of how contextual information influences children's moral judgment about third-party interventions and shed light on the developmental trajectory of children's sociomoral cognition.

Lay Summary

Punishment is theorized to be essential for human cooperation. Studies have shown that preverbal infants selectively prefer agents who hinder rather than help previously antisocial individuals. However, preschoolers uniformly prefer helpers over hinderers without considering the target's previous prosocial or antisocial actions. One possible reason is that children are so averse to antisocial acts that they endorse punishment only in highly specific contexts, such as when performed by authority figures like police or teachers. To explore this hypothesis, the present study investigated whether children are selective in their evaluations of third-party interventions when the punishing agents are depicted as authority figures (e.g., police officers). Contrary to our prediction, both 3- and 4-year-old children negatively evaluated the punishing police officers and positively evaluated the rewarding police officer regardless of whether the target was previously prosocial or antisocial. These findings shed light on the developmental trajectory of children's sociomoral cognition.

Preface

This thesis represents the unique and unpublished intellectual creation of Y. Ma. The author conducted the research at the University of British Columbia, with guidance and supervision from J. K. Hamlin, who played a role in the research design, data analysis, and research process. The University of British Columbia's Research Ethics Board granted approval for all aspects of the work, including data collection procedures and methodologies (Approval: H10-01808, titled "Early Understanding of the Physical and Social Worlds").

Table of Contents

List of Tables

List of Figures

Acknowledgements

Firstly, I would like to express my heartfelt gratitude to Dr. J. Kiley Hamlin for her invaluable guidance during my time at the University of British Columbia, as well as for her mentorship and instruction in conducting this research. Our discussions on research questions and conceptualizing projects were truly enlightening and enriched my learning experience.

I am also deeply thankful to Dr. Andrew Baron and Dr. Susan Birch for their roles as my committee members. Additionally, I extend my appreciation to the other esteemed faculty members in the Department of Psychology for their unwavering support throughout my studies at University of British Columbia.

My heartfelt thanks go out to the members of the Center for Infant Cognition: graduate students Rachel Drew, Francis Yuen, and Zohreh Soleimani, for their companionship and support on this journey. I am grateful for the assistance of former lab manager Fibha Khan and Sydney Lopes, as well as the current lab manager Natalia Modzelik and Chole Fichter, for their kind help. Special thanks to undergraduate research assistant Hattie Zhang for her invaluable assistance in data collection.

I cannot overlook the tremendous support and encouragement I received from my friends Dr. Xin Sun, Huixian Yu and Chuyan Qu during the thesis writing process. I am also grateful to Yuan Yao for his support during challenging times. Special thanks to my cat friends Guapi, Java and Taiji.

Last but not least, I extend my heartfelt thanks to all the children and families who willingly participated in our study. Their involvement was crucial in contributing to the success of this research.

Introduction

Throughout the evolution of human society, humans rely heavily on cooperation in food gathering, hunting, and protecting the tribe from enemies, making it one of the most crucial factors in the survival and success of humans (Fehr & Gächter, 2002; Fessler & Haley, 2003). Nowadays, most social activities require cooperation to properly function. Although cooperation is often costly (e.g., time, effort, money, etc.), every individual of the "team" would benefit from the team's success. However, even those who failed to contribute to the costly cooperative work would receive benefits. Without deterring potential free-riding, people will be less likely to keep participating in costly cooperation, which would jeopardize the cooperation system. So how could cooperation evolve in large-scale societies and what mechanism makes stable cooperation persist in human society?

Cooperation and Punishment

So far there have been several theories that explored the answer to this problem. Kin selection theory looked at cooperation through the nature selection lens. It proposed that human altruism would occur when the actor's genetic relatedness with the recipient, multiplied by the benefit, exceeded the actor's cost, and therefore help the shared genes to increase in frequency (Hamilton, 1964). However, kin selection theory failed to explain the pervasive costly cooperation among genetic-unrelated individuals. The indirect reciprocity theory believed the shared moral systems and individual reputation are the base stone of cooperation. According to the indirect reciprocity theory, people's costly prosocial behavior could serve as an investment to build good reputation, and thus increase the actor's chance to receive benefit in the future

interactions (Hilbe, Schmid, Tkadlec, & Nowak, 2018; Nowak & Sigmund, 1998). Likewise, the costly signalling theory believes that by cooperating and contributing to the group's welfare, the signaler could advertise their quality as reliable allies, mates, or competitors, and thus receive future payoff (e.g., being chosen as mates or allies) (Gintis, Smith, & Bowles, 2001).Yet both the indirect reciprocity theory and the costly signalling theory are based on the assumptions that all community member share a consensus on what is good and bad, and each individual's reputation is publicly salient, and therefore can not explain people's altruism in non-repeated interaction among different moral dyads, when reputational benefit is absent (Fehr & Gächter, 2002).

Punishment, however, provides an explanation for the costly altruistic behavior among unrelated members in non-repeated interactions. Just as Machiavelli said in *The Prince*, "it is much safer to be feared than loved... for love is preserved by the link of obligation which, owing to the baseness of men, is broken at every opportunity for their advantage; but fear preserves you by a dread of punishment which never fails." Indeed, studies found that if those who cooperate could impose a costly punishment on those who violated cooperative norms, sustainable cooperation can develop among non-related individuals in anonymous interaction (Boyd, Gintis, Bowles, & Richerson, 2003; Gardner & West, 2004; Yamagishi, 1986). Through adding cost to uncooperative behaviors, punishment reduces the likelihood of the defector's repeating transgression. Evidence demonstrated that punishment towards uncooperative others is common across human societies (Bernhard, Martin, & Warneken, 2020; Bull & Rice, 1991; Henrich et al., 2006; Sober & Wilson, 1998; Thaler, 1988; Yamagishi, 1986) and even among social animals (Clutton-Brock & Parker, 1995; Hauser, 1992). For example, rhesus macaques (a species of monkey) that locate food sources but refrain from alerting others through food calls are at a higher risk of facing aggression, compared to those individuals who do announce their findings

through food calls (Hauser, 1992). Another study looked at people's reactions in the "Public Goods" game scenario, which featured each participant voluntarily choosing the number of resources they would contribute to the public investment pool, and the payback was shared equally among all participants. The result showed that after the failure of voluntarily based cooperation, instead of trying to induce free-riders into mutual cooperation through cooperative actions, members spontaneously developed a negative sanctioning system and successfully assured stable cooperation (Yamagishi, 1986). Additionally, empirical evidence from across 15 diverse populations has revealed a positive correlation between costly punishment and altruistic behavior (Henrich et al., 2010). To summarize, by creating an expectation that the violation of norms would be sanctioned, punishment serves to deter free-riding and foster cooperation in social groups (Fessler & Haley, 2003; Fehr & Fischbacher, 2004).

Third-party punishment, where the punisher's interest is not directly affected by the transgression, is crucial to maintaining an effective punitive system. Firstly, in large-scale societies, the chance of repeated interaction between the same moral dyads is rather low compared to non-repeated interaction with unknown individuals (Henrich et al., 2010). If we exclusively rely on the victim's direct punishment, a rather limited number of norm perpetrators would be sanctioned. Secondly, there could be no direct victim in many norm transgression cases (e.g., if a soldier flees from a battle, the desertion would cost little harm on any particular individual), and third-party punishment could significantly broaden the reach of norm enforcement in situations like these (Fehr & Fischbacher, 2004). Therefore, it is important to rely on punishment which is enforced by third-parties whose economic payoff is not harmed by the norm violation (“third-party punishment”) (Bendor & Swistak, 2001; Fehr & Fischbacher, 2004).

Evidence has shown that people across human cultures and societies are willing to punish, even at personal cost, when acting as unaffected third-party (Fehr & Gächter, 2002; Henrich et al., 2010; Raihani & McAuliffe, 2012). But why, after all, should we voluntarily punish people not because of what they did to us, but because of what they did to others? What incentive motivates people to willingly sacrifice their own interests to punish, rather than passively observing without intervening? If punishment and its benefits are shared equally among group members, sanctioning individuals who take the risk and pay the cost to punish are at disadvantage compared to people who cooperate but do not punish, which eventually undermines the third-party punishment system (Barclay, 2006). Turns out, the punishment toward norm violation are themselves a form of second-order public good (Fessler & Haley, 2003; Vaish et al., 2016), and norm enforcement can be considered a second-order cooperative behavior since all group members could benefit from the deterrence of norm violation (Barclay, 2006; Wedekind & Milinski, 2000; Yamagishi, 1986). Just like first-order cooperation, third-party punishment could be encouraged and maintained if punishers can receive more benefit than other group members for their cooperative punitive action (Barclay, 2006).

Indeed, evidence showed that third-party punishers would receive more benefits than the non-punishers for punishing unfair, uncooperative behaviors (Barclay, 2006). Studies found that third-party punishers would receive more preference from others (Gordon, Madden, & Lea, 2014), were evaluated as more worthy of trust (Nelissen, 2008) and respect (Barclay, 2006), were more likely to be chosen as partners or mates (Gintis, Smith, & Bowles, 2001; Nelissen, 2008), and tend to receive more material rewards from others (Wedekind & Milinski, 2000). Furthermore, since third-party punishment constitutes a cooperative action for second-order public good, people sometimes enforce higher-order punishment toward those who failed to

punish when they should have, which suggests that third-party punishment could be viewed as not only preferable, but also normative in some contexts (Martin et al., 2019).

The development of altruistic punishment

As discussed in the previous section, humans have built and maintained a complex punitive system to sustain cooperation, empathy, and altruistic behaviors in societies. However, an important question arises: at what point does the punitive system used to support third-party punishment become functional in ontogeny? In other words, when do children develop the sophisticated practice of rewarding punishers and punishment itself in order to uphold dependable norm enforcement and cooperation?

To understand the essence of altruistic punishment, children must first possess a moral sense, which encompasses the fundamental concepts of good and bad, right and wrong, and the deservingness of rewards and punishment. It is argued that three crucial moral capacities are relevant to comprehending cooperation and punishment: 1) moral goodness, which involves feeling empathy and helping others regardless of personal cost; 2) moral evaluation, which entails assessing and disapproving of others' uncooperative, unempathetic, or unhelpful social behaviors; and 3) moral retribution, which involves enforcing or positively evaluating punishment for antisocial actions (Hamlin, 2013). With a developed moral sense, children gain the ability to accurately analyze others' cooperative or uncooperative, helpful or unhelpful behavior, assess and interact with them based on these observations, and enforce and endorse punishment. And this leads to the following two questions: 1) How do children analyze, evaluate, and respond to others' antisocial actions as third parties throughout development? 2) How do children evaluate individuals who stand out and act as third-party punishers? Would

children take the specific context into consideration when they evaluate punishment? Will they adore them because they sanction wrongdoers, or will they dislike them because they engage in antisocial behavior?

How do infants and toddlers evaluate and react to antisocial others?

Starting at their earliest years, young children have experience with third-party punishment such as "time-out" or removal of toys when transgression happens. A large body of research has explored young children's moral evaluation and response to antisocial events and antisocial others as third-parties. For example, one study presented 6- and 10-month-old infants a goal-reaching scenario where the protagonist is trying to reach the top of a hill, with the helper agent pushing them up the hill and the hinderer agent pushing them down. The result revealed that preverbal infants were able to identify an moral agent's helpful/unhelpful action toward a third-party individual, and selectively prefer the helper over the hinderer (Hamlin, Wynn, & Bloom, 2007; see also Scola, Holvoet, Arciszewski, & Picard, 2015). Similar studies tested infants with different goals in various social scenarios (e.g., opening a box to get a toy or retrieve a dropped ball) repeated the result of this study and demonstrated that preverbal infants robustly prefer prosocial others over antisocial others (Hamlin & Wynn, 2011). For infants as young as 3-month-olds (who are not capable of reaching for objects), researchers used similar stimuli as in Hamlin, Wynn, & Bloom (2007) and tested with the preferential looking method. They found that infants selectively preferred to look at the agent who helped the climber over the agent who pushed the climber down the hill, which arguably indicated that infants preferred and would like to interact with the prosocial character over the antisocial character (Hamlin, Wynn, & Bloom, 2010). Another study found that 9-month-old infants prefer agents who attempt but were unable to give them a toy (e.g., accidentally dropping the toy), compared to agents who intentionally

withheld the toy (e.g., playing with the toy themselves or teasing the infant with the toy) (Behne et al., 2005). These results demonstrate that preverbal infants and toddlers are capable of distinguishing others' nice or mean intentions, identify moral transgressions accordingly, and appropriately evaluate prosocial/antisocial others based on their action.

How would young children react to unhelpful/uncooperative/unempathetic others? A study revealed that at around 21-month-old, young children tend to give more treats to the previously prosocial puppet, and take more treats from the previously antisocial puppet (Hamlin et al., 2011). Another study found that 2-year-old children spontaneously protested (i.e., verbal correction or correction through physical demonstration) when their own interests were harmed by a transgression (i.e., when a puppet threw away the child's hat), but they did not protest for actions that did not constitute a violation while having the same low-level physical features (i.e., when a puppet threw away their own hat). This showed that by at least 2 years of age, young children were able to understand and react to moral violations. The same study also found that it was not until 3 years old that children protested as a third-party when another individual was the victim of the transgression (i.e., when a puppet threw away the hat of a third party) (Rossano, Rakoczy, & Tomasello, 2011), which showed that at latest at 3 years of age, young children began to sanction antisocial behaviors as third-parties. Furthermore, 18- and 25-month-old would sympathize with and direct prosocial behavior to victims of antisocial others, even without the victim's overt emotional cue (Vaish, Carpenter, & Tomasello, 2009). To sum up, it seems that young children are able to analyse others' prosocial/antisocial actions, and interact with the prosocial/antisocial others appropriately.

How do infants and toddlers evaluate those who punish?

Punishment, despite being overtly antisocial, is appropriate if targeted at wrongdoers and serves a prosocial end. Therefore, the capacity to analyze and positively evaluate altruistic punishment requires young children to be able to assess an action not solely based on its immediate negative valence, but also based on the target to which it is directed.

Studies have looked at the developmental trajectory of young children's evaluation of punishment and have found that infants seemed to be sensitive to the target's previous prosocial or antisocial action in their moral evaluation. They positively evaluate punishment when it was directed towards a previously antisocial individual, before they were able to actively enforce punishment towards antisocial others on their own. For instance, while 5-month-old infants preferred prosocial individuals regardless of whether the action was directed towards a helpful or unhelpful target, 8-month-old infants would take the status of the target into consideration in moral evaluation. When forced to choose, they selectively preferred those who were prosocial to helpful others and selectively preferred individuals who were antisocial to unhelpful others (Hamlin et al., 2011). Another study presented infants with a scenario where a character climbed a hill, and the helper pushed the climber up the hill while a hinderer pushed the climber down the hill. The result revealed that when the climber's gaze was consistent with the goal of reaching the top of the hill, infants selectively preferred the helper over the hinderer; however, when the climber's gaze was inconsistent with the goal of reaching the top of the hill, infants showed no preference between the characters (Hamlin, 2015). This finding indicated that infants were able to selectively make moral judgement based on the context in which helping and hindering occur. A relevant study also found that 21-month-olds would expect a third-party puppet to enforce physical punishment (i.e., hitting with a stick or pushing strongly) toward an agent who refused

to defend his partner from an aggressor (Geraci, 2021). Such pattern aligns with the results that young children, as third parties, would act nicely towards prosocial individuals and act negatively towards antisocial individuals (Hamlin et al., 2011; Rossano, Rakoczy, & Tomasello, 2011; Vaish, Carpenter, & Tomasello, 2009).

To summarize, young children would enforce and positively evaluate those who enforce third-party punishment, which suggests that they might consider the target's previous action when judging punishment and see antisocial actions as deserving to be treated negatively. Additionally, the fact that they would view a locally negative action as positive when it is directed towards antisocial others suggests that infants were not limited to the most immediate prosocial/antisocial valence of an action, but were also capable of considering the specific context of an action and analyzing its global valence in their moral evaluation (Geraci & Surian, 2011). That is, they were capable of understanding an act based not merely on its own nice or mean value, but also on its target (Hamlin, Wynn, Bloom, & Mahajan, 2011; Li & Tomasello, 2018).

Preschooler's evaluation of third-party punishment

As for older children, many empirical studies have investigated preschool-age children's engagement in third-party punishment toward defectors. For example, children as young as 3-year-old would spontaneously protest against antisocial others (Rossano, Rakoczy, & Tomasello, 2011). 4-year-old children distributed more resources to a previous helper than the hinderer (Kenward & Dahl, 2011). 5- and 6-year-olds would prevent the unfair individual from receiving resources at a cost (McAuliffe, Jordan, & Warneken, 2015). 5-year-old children would allocate unpleasant items to antisocial adults (Kenward & Östh, 2015). Additionally, starting

from the age of 3-year-old, children would intervene and punish even if it required sacrificing their own resources, and the rate of costly punishment would increase with age (Yudkin, Van Bavel, & Rhodes, 2020).

However, some of the evidence from older children indicates that they experienced a conflict between their punitive motivation and the desire to act prosocially. For example, in Kenward & Dahl's (2011) study, although 4-year-old would distribute less resource to the hinderer when the resource was scarce (i.e., 3 biscuits), they would distribute equally to the helper and the hinderer when there was plenty of resource (i.e., 8 or 9 biscuits). Specifically, even in the odd-plenty number of resource conditions (9 biscuits), the majority of the 4-year-old would reach an equal distribution by not distributing all the biscuits. This result indicated that 4-year-olds were reluctant for unequal distribution, even when the potential victim was previously antisocial and the potential beneficent was previously prosocial. As a comparison, 3-year-olds in the same study would uniformly share more resources to the helper no matter if the resource is scarce or not. Based on such evidence, we may infer that with development, children become more sensitive to the punitive approach and less willing to engage in third-party punishment in an antisocial manner.

Meanwhile, an increasing number of studies have examined children's evaluation of others' third-party intervention. Some evidence demonstrated that children would positively evaluate punishment and punishers. For example, a study found that children aged 5 to 9 positively evaluated third-party punishers who sacrificed their own belongings to sanction an unfair allocator (Lee & Warneken, 2020). In another study, 5-year-old children watched a video depicting a transgressor intentionally breaking the victim's belongings. In their video, an enforcer verbally accused the transgressor, while a non-enforcer just made neutral comments. The results

revealed that 5-year-olds, but not 4-year-olds, preferred enforcers, evaluated enforcers more positively, and distributed more resources to them (Vaish et al., 2016). To note, in the similar context of violation of property rights, it was observed that even 3-year-old children spontaneously protest verbally as third parties (Vaish et al., 2011). The comparison between these results suggest that the ability to value third-party punishment may emerge later in development than the capacity to enforce their own punishment.

Another study found that 6- to 11-year-old students approved of punishment through reporting a peer's transgression to an authority figure (i.e., a teacher). Specifically, older children deemed reporting appropriate only for major transgressions, whereas younger children did not consider the seriousness of the transgression when evaluating the appropriateness of reporting (Loke et al., 2011). It is also revealed that although 5- to 9-year-old children would positively evaluate third-party punishers in unfair resource allocation, they preferred helpers who helped the victim over punishers who sanctioned the transgressor (Lee & Warneken, 2020).

However, there is also opposing evidence that preschool-age children do not approve of third-party punishment in some contexts. In one study, researchers presented 3- and 5-year-old children with puppet show videos depicting third-party rewarding and punishment scenarios. The children were asked to rate the rewarder and punisher in each scenario. The first video showed a prosocial agent sharing food with others or an antisocial agent stealing a toy from others. The second video featured a helper attempting to assist the previous prosocial or antisocial target in opening a box, or a hinderer attempting to close the prosocial or antisocial target's box. Based on the outcome of the helping or hindering actions, there were four different scenarios: the successful hinderer, the failed hinderer, the successful helper, and the failed helper. The children were then asked to rate the actions of the puppet using a moral scale. The results revealed that

although 3-year-olds positively rated punishers who intended to punish but failed to do so (i.e., tried to prevent the puppet from opening the box but failed), they negatively evaluated those who successfully punished individuals who stole toys from their peers (i.e., successfully prevented the puppet from opening the box). On the other hand, the target's previous antisocial action did not seem to influence the moral evaluation of 5-year-old children, as they negatively evaluated both successful and unsuccessful third-party punishers (Li & Tomasello, 2018).

A recent study investigated whether 3- and 4-year-olds would consider contextual information in moral evaluation. The researchers used animations to explore if the social and moral evaluation of 3- and 4-year-old children towards a helper or hinderer would differ depending on whether it was directed towards a previously helpful or unhelpful target. In the first show, a star attempted to build a tower, where the helper assisted the star in completing the tower, and the hinderer broke the tower. In the second show, the helper or hinderer from the first show played with a ball but accidentally dropped it. The rewarder then returned the ball, while the punisher took it away. After watching the shows, the children were asked to select between the rewarder and the punisher based on liking, friendship, and rightness. The results showed that both 3- and 4-year-olds consistently evaluated the rewarder positively and the punisher negatively, regardless of whether the action was targeted at a previously prosocial or antisocial individual (Van de Vondervoort, 2020). This suggests that children aged 3 and 4 may be insensitive to contextual information.

Another study used video-recorded puppet shows and investigated whether children aged 2 to 7 years old would consider the context when evaluating helpful and unhelpful agents, depending on whether the agent was being asked to help in a morally good action (i.e., asking for help to assist a third-party in building a tower) or a morally bad action (i.e., breaking someone's

tower). The study found that 2- to 4-year-old children would always evaluate helping as morally good and not helping as morally bad, even when the goal of the recipient was morally bad. Only 4.5- to 7-year-old children rated helping to complete an immoral act as immoral and not helping in an immoral act as moral. The results demonstrated that younger children aged 2 to 4 years old were not sensitive to contextual information, as they positively evaluated the helper regardless of the goal of their helping action. However, older children over 5 years of age preferred agents who refused to help in an immoral act compared to those who helped (Myslińska Szarek, Baryła, & Wojciszke, 2023).

To sum up, contrary to younger children, children over 5 years of age seem to be capable of going beyond the local valence of an action and consider the target's previous prosocial or antisocial behaviors when evaluating altruistic punishment (Lee & Warneken, 2020; Loke et al., 2011; Vaish et al., 2016). However, children between 3 and 5 years of age may not be sensitive to contextual information (Li & Tomasello, 2018; Li et al., 2020; Van de Vondervoort, 2020). Furthermore, compared to younger children, older children tended to endorse only milder punishment in limited contexts and preferred using a prosocial approach, such as compensating victims, when faced with a choice between punishment and a prosocial action (Lee & Warneken, 2020; Loke et al., 2011). To summarize, it is currently unclear whether 3- to 5-year-old children are able to incorporate contextual information into their moral evaluations of punishment. Furthermore, it remains unknown in which specific contexts they perceive altruistic punishment as appropriate.

Authority figures as Punishers

It is argued that authority hierarchy serves as a fundamental framework for human social interactions and entails moral responsibilities for both authority figures and their followers. People tend to believe that authority figures have a moral duty to lead, guide, direct, and safeguard their followers, while followers have a moral obligation to demonstrate respect, obedience, and deference towards their authority figures (Rai & Fiske, 2011). Previous evidence demonstrated that infants, children, and adults perceive authorities as being more obligated to punish (Marshall, Mermin-Bunnell, & Bloom, 2020; Martin et al., 2019; Stavans & Baillargeon, 2019). Therefore, preschool-age children's disapproval of others' third-party punishment revealed in past studies (Li & Tomasello, 2018; Li et al., 2020; Van de Vondervoort, 2020) might be because of their lack of expectation and endorsement of the responsibility for punishment in a non-authority individual.

In fact, peer punishment of norm violation is often discouraged in modern society, since such punishment is often viewed as aggressive and, therefore, more problematic than beneficial to the social group (Eriksson, Andersson, & Strimling, 2016). Imagine when someone has committed a crime, it is generally perceived as more legitimate for a judge to sentence the criminal to prison. However, it would be less appropriate for an ordinary individual to imprison the wrongdoer in their own home. The distinction lies in the fact that the former involves punishment enforced by a legally trained and institutionalized authority figure, whereas the latter involves punishment enforced by a lay individual who lacks both the power and legal knowledge. As a result, the punishment imposed by a legal authority figure is generally perceived as having more legitimacy compared to the punishment imposed by a lay third party.

Indeed, studies have found that compared to non-authority peers, people judged authority figures as more obligated to punish transgressors (Martin et al., 2019). Developmental studies have demonstrated that 17-month-old infants expect leaders to intervene in transgressions and rectify wrongdoing, but they have no expectation for intervention from a non-leader individual (Stavans & Baillargeon, 2019). Similarly, 4- to 7-year-old children judged authority figures as obligated to intervene and punish in transgressions (Marshall, Mermin-Bunnell, & Bloom, 2020). Taken together, we might induce that preschool age children may consider third-party punishment as role-dependent, and judge punishing an antisocial target as appropriate when the punisher holds authority status.

The present study

Building on the previous work, the current study aimed to address this unsolved question: Are 3- and 4-year-olds really insensitive to any contextual information in their moral evaluation of helping and harming behaviors, always positively evaluating helping and negatively evaluating punishment? Or do they positively evaluate third-party punishers in some, albeit limited cases? Specifically, is it okay for authority figures, who are obligated to enforce social norms, to punish wrongdoers?

We used live puppet shows to present children with scenarios in which an authority figure (i.e., a police officer puppet) rewarded or punished a target who previously helped or hindered others. Children were asked to make force-choice selections between the rewarder and punisher based on: 1) their liking, 2) their perception of rightness, and 3) their assessment of the police officer as a good authority figure. Additionally, they were asked to rate the appropriateness of the authority figure's rewarding or punishing actions using a 5-point Likert scale.

It was predicted that both 3- and 4-year-old children would select the punishing police officer over the rewarding police officer in their social ("liking") and moral ("rightness", "good police officer") preference when the target that were rewarded or punished were previously antisocial, and select the rewarding police officer over the punishing police officer when the target that were rewarded or punished were previously prosocial. We also predicted that both 3- and 4-year-old children would evaluate the punisher more positively when the target that was rewarded or punished were previously antisocial than when they were previously prosocial; and children would evaluate the rewarding police officer more positively in the prosocial target condition than in the antisocial target condition.

Method

Participants and Design

Participants were twenty-eight 3-year-olds (16 girls; mean age 3;6; range 3;0,11 - 3;11,28) and thirty-five 4-year-olds (13 girls; mean age 4; 6; range = 4;0,8 - 4;11,28). A 2 (age: 3 years old vs. 4 years old; between-subjects) × 2 (target: prosocial versus antisocial; between-subjects) × 2 (police action: punish versus reward; within-subjects) mixed design was applied to test the hypotheses. Specifically, sixteen 3-year-olds (9 girls; mean age 3;7; range = 3;0,11 - 3;11,28) and nineteen 4-year-olds (8 girls; mean age 4;6; range = 4;0,17 - 4;10,17) were assigned to the prosocial target condition, while twelve 3-year-olds (7 girls; mean age 3;6; range = 3;1,28 - 3;11,16) and sixteen 4-year-olds (5 girls; mean age 4;6; range = 4;0,8 - 4;11,28) were assigned to the antisocial target condition. Nine additional 3-year-olds and four 4-year-olds were tested but excluded from the final analysis due to failure of passing screening questions (8), unwillingness to respond to questions (2), procedural errors (2) and parental inference(1).

Procedure

Warm-up. Children were taken to a quiet test room and asked to sit by a table. An experimenter sat on the other side of the table facing the children. Children were first shown a picture of a playground, and asked to 1) point to characters on the swings and stairs, to warm up a pointing response; 2) verbally identify the color of the stairs, to warm up a verbal response.

Scale familiarization task. After the warm-up, children completed a scale familiarization task to teach them how to use the moral rating scale, a 5-point likert scale to rate the acceptability of an action. The scale ranged from 1 (“really bad”, with two thumbs down and a

very sad face), 2 (“a little bit bad”, with one thumb down and a slightly sad face), 3 (“just ok”, with a sideway thumb with a neutral face), 4 (“a little bit good”, with one thumb up and a small smiley face), to 5 (“really good”, with two thumbs up and a big smiley face). After introducing the scale, children practiced by rating various scenarios (e.g., “What do you think if someone shares a birthday cake with their friends?”) and were asked to point to the corresponding picture in the scale. Lastly, to check if children understood and remembered the scale, they were instructed to identify points on the scale (e.g., “where is the picture for really bad things?”). The experimenter corrected them if they pointed to/indicated the wrong picture and asked again. Children who were unable to correctly answer the final scale questions were excluded from the final sample.

Box show. This show was adapted from previous infant studies (e.g., Hamlin & Wynn, 2011) and a preschool-age child study (Van de Vondervoort & Hamlin, 2017). Children watched a live puppet show involving 5 characters: a protagonist who wants to open a box containing a toy, a helper who helps the protagonist to open the box, a hinderer who closes the box, and two police officers who witness the helping/hindering behavior. To differentiate the helper/hinderer and the two police officers, the characters wore t-shirts of different colors (i.e., green/red for helper/hinderer and orange/purple for police officers). The t-shirt colors of the helper/hinder puppets, the order of the helper/hinder’s actions, and the side of the stage of the helper/hinderer were counterbalanced.

The experimenter started by introducing the helper/hinderer (bunnies), and pointed out the color of the characters’ t-shirt (i.e., “This guy wears a green shirt, and this guy wears a red shirt!”). After the introduction the helper or the hinderer puppet were each placed at either the left or right corner of the table respectively. Then the two police officer characters (puppies)

were introduced as police officers and wore police uniforms (i.e., “We also have two police officers. Look! We know they are police officers because they have their police uniforms on!”). The experimenter then pointed out the color of the police officers’ t-shirt (i.e., “This police officer wears a purple shirt, and this police officer wears an orange shirt!”). After introducing the police officers, the experimenter checked if children understand the role of police officers (i.e., “Do you know what police officers do?”, “That’s right, they make sure everyone follows the rules!”). Then both the police officers were placed at the side of the table on the children’s right. Then the experimenter introduced the protagonist (duck) (“And we also have a duckie!”), and placed a clear plastic box containing a toy at the center of the table.

At the start of each trial, the protagonist entered from the center of the back of the table, waved to the children and said: “Hi!”, and ran to one side of the box. He turned his head and looked into the box twice and said:“Look! A toy!”. Then the protagonist jumped on the front corner of the top of the box, and attempted to open the box to reach the toy four times. During the initial two attempts, the protagonist pulled up, lifted the edge of the box a few inches while shaking slightly as if the lid was too heavy to hold on to, and then dropped it back down. On the third and fourth tries, he lifted the edge of the lid and lowered it while continuously grabbing the lid and saying: “So heavy!”. On the fifth attempt, the helper/hinderer who sat on the other side of the table from the protagonist ran forward and intervened.

The intervention was either prosocial or antisocial. During the prosocial intervention event, the helper puppet on the opposite side of the stage from the protagonist ran forward, jumped up and grabbed the front corner of the box, and opened the box together with the protagonist while saying “Open!”, helping the protagonist get its goal. After the box was opened, the protagonist moved into the box and grabbed the toy, as if he was happy to be able to play

with it. The helper then jumped off the box and ran backward off the table, and finally the protagonist sat up and also ran off the table. During the antisocial event, the hinderer puppet on the opposite side of the stage from the protagonist ran forward, jumped up, and sat down on top of the box, slamming the box shut while saying: “Close!”, preventing the protagonist from getting its goal. The protagonist then jumped off the box and lay down beside it, as if he was sad and disappointed for not getting the toy. The hinderer then ran backward off the table, and finally the protagonist sat up and also ran off the table. Each of the prosocial/antisocial intervention events was played once, and the sequence was counterbalanced. All of the characters’ narratives in the show were performed with a high-pitched voice, so as to indicate they were spoken by the puppet rather than the experimenter.

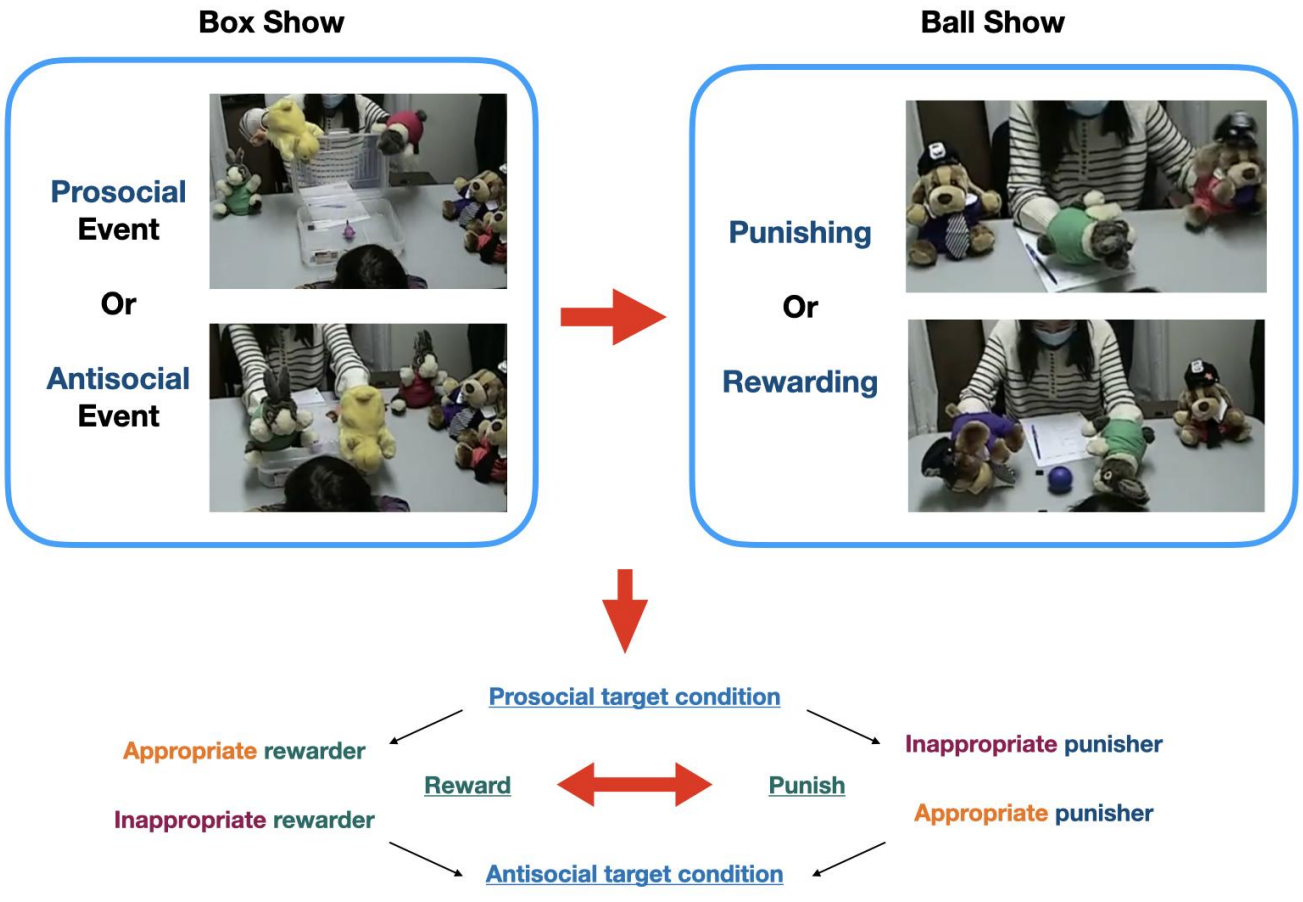

Fig. 1. The procedure of the experiment. Children first watched two *Box shows* involving the prosocial puppet helped and the antisocial puppet hindered the protagonist (duck) when he tried to open the box, with the police officers (puppies) watching on the side. Then children watched the *Ball show* showing the interactions between the police officer puppet and either the prosocial or the antisocial puppet. One of the police officers gave the ball back (rewarder), another police officer took the ball away (punisher).

After the box shows, the experimenter took the box away and placed the bunnies on the same side of the stage as they had during the box show. A series of manipulation check questions were asked to assess children's understanding, retention, and evaluation of the puppet shows. Children answered questions in the following sequence:

1)"*Who closed the box? Who opened the box?*".

2) "*Did the police officers here see this guy (point to the helper) open ducky's box and this guy (point to the hinderer) close ducky's box?*".

3) "*Who do you like?*"

4) Children were asked to point to the red/green puppets in the opposite sequence from the previous questions: *"Can you point to the green guy", "Can you point to the red guy?"* The first puppet to be pointed to for this question should be different with the first puppies being pointed at the previous questions. This question served to avoid noise from maintaining the same sequence of pointing at puppets.

5) "*Who was nicer?*"

6) "You know what, I think one of these guys should get in trouble. *Who do you think should get in trouble?*"

To encourage children to answer questions and to facilitate memorization for the manipulation check questions (questions 1, 2, 5, or 6), the experimenter repeated the answer after children correctly answered the question (e.g., “That’s right! He did open/close it!”, “That's right, they were watching the whole time!”). If children incorrectly answer any of the manipulation check questions (questions 1, 2, 5, or 6), the experimenter repeated both the prosocial and antisocial intervention events, and asked all of the questions again. If children answered manipulation check questions wrong for the second time, they were removed from the final sample. No reinforcement was given for the liking question after children’s answers.

Ball show. After watching the box show and correctly answering the comprehension and memory check questions, children watch a live puppet show involving 3 characters from the previous box show: the helper or the hinderer (who became the protagonist of the ball show in the prosocial target/antisocial target conditions, respectively), the rewarder police officer, and the punisher police officer. Children were randomly assigned to either the prosocial target or the antisocial target condition.

At the beginning of the show, the experimenter put a blue rubber ball at the center of the table, and placed the two police officers at the left and right back corner of the table respectively (position of the punisher/rewarder counterbalanced), and sat the helper or hinderer from the box show at the center of the back of the table. The experimenter said: “Now we're going to see a new show with this guy (pointing to the helper/hinderer puppet) and these police officers (pointing to the police officer puppets)”. Before each trial of the ball show, the experimenter checked if children remembered the helper/hinderer in the previous show (i.e., ”Did this guy open the box or did he close the box?”, “That's right! He did opened/closed it!”) and if children understand that the police officers had witnessed the helping/hinderer intervention (e.g., “And

did the police officers see him open the box? That's right, they were watching the whole time!"). If children were unable to correctly answer the check questions, the experimenter would correct them (e.g., "Actually this guy opened/closed the box.").

The protagonist moved from the center of the back of the stage, grabbed the ball, then jumped up and down and repeatedly tossed and caught the ball three times. When he tossed the ball for the fourth time, he lost the ball to one side of the table and said: "Whoops!". Then the police officer puppet who sat at the corner where the ball was lost ran forward and retrieved the ball. The protagonist then turned to the police officer and asked for the ball back, saying: "Hi! Can I have it?" The police officer turned and looked at the protagonist in response, and then both puppets turned forward again. This sequence was repeated once more, and finally on the protagonist's third turn the events diverged. During the *reward* event, the police officer leaned down and rolled the ball back to the protagonist while saying: "Here!", and the protagonist caught the ball. The police officer then ran off stage, and the protagonist holding the ball jumped twice and then ran off the table. In the *punishing* event, the police officer holding the ball ran away to the back of the table while saying: "Bye!". The protagonist turned his head to watch the police officer run away, and then faced forward without the ball, jumped twice and then ran off the table. The rewarding and the punishing event were each played once, in counterbalanced order.

After the show, children were presented with both the rewarding and punishing police officers (who sat on the same side of the stage as they had during the show), and were asked manipulation check questions (i.e., "Who took the ball away? Who gave the ball back?"). If children were unable to correctly answer these questions, the researcher would repeat the show. Following the check questions, children were asked three **test questions** in the following order:

1) *"Who do you like? Why do you like him?*"

2) "*Who do you think did the right thing? Why do you think he did the right thing?*"

3) "*Who do you think is a good police officer? Why do you think he is a good police officer?*".

Note that before test question 2, children were asked to point to each of the police officers (e.g., "Can you point to the purple police officer? Right! Can you point to the orange police officer? Right!"), in the opposite order as during the liking question, to prevent perseveration.

After the first round of puppet shows/questions, children were shown both the rewarding and punishing events a second time, and were presented with n the likert scale, After each event, children were asked to point to the scale to **rate the rewarding/punishing action** of the police officer in the event (i.e., "*What do you think of what he did (pointing to one of the police officers*)?".

Coding Procedure

Children's responses regarding their reasons for choosing the selected puppet in the three test questions were recorded and coded into the following categories, adapted from Van de Vondervoort's (2020) coding scheme:

1) *Uninformative responses*, which included i) meaningless, unintelligible answers; ii) answers irrelevant to the puppet show (e.g., "Because I like purple."); iii) uninformative answers (e.g., "I don't know."); iv) cases when no verbal response was provided.
2) Response on *moral actions and motivations*, which included answers focused on the agent's moral actions (e.g., "Because he gave the ball back.") and the evaluative moral attributes of the agent (e.g., "Because he is nice.").

3) Response on *social actions and motivations,* which included descriptive, non-morally evaluative answers about the action (e.g., “Because he wanted to.”) and answers about social motivations behind the action (e.g., “Because he is happy.”).
4) Response on *target’s previous prosocial/antisocial action,* which referred to answers centered on the target’s previous helping/hindering action (e.g., “Because the bunny helped.”).

Results

Forced-choice judgments. We conducted binomial tests for the three forced-choice questions (i.e., liking, rightness, good police officer) to examine if children's likelihood of selecting the rewarder/punisher would differ from chance level.

As shown in Figure 2, in both the prosocial target and antisocial conditions, 3-year-olds' selection for the rewarder/punisher were not significantly different from chance levels when asked about liking (prosocial target: 10/15 selected the helper, $p = .151$; antisocial target: 6/13 selected the helper, $p = .500$). However, contrary to the hypothesis, 3-year-olds were more likely to select the rewarder in both the prosocial target and antisocial conditions when asked to select the police officer who did the right thing (prosocial target: 14/15 selected the rewarder, $p < .001$; antisocial target: 11/13 selected the rewarder, $p = .013$), and identify the rewarder as the good police officer (prosocial target: 14/15 selected the rewarder, $p < .001$; antisocial target: 10/13 selected the rewarder, $p = .048$).

Likewise, regardless of the prosocial target or antisocial condition, 4-year-old children preferred to select the rewarder police officer when asked about liking (prosocial target: 17/19 selected the rewarder, $p < .001$; antisocial target: 14/16 selected the rewarder, $p = .003$), rightness (prosocial target: 16/19 selected the rewarder, $p = .003$; antisocial target: 15/16 selected the rewarder, $p < .001$), and chose the rewarder as the good police officer (prosocial target: 17/19 selected the rewarder, $p < 0.001$; antisocial target: 15/16 selected the rewarder, $p < 0.001$).

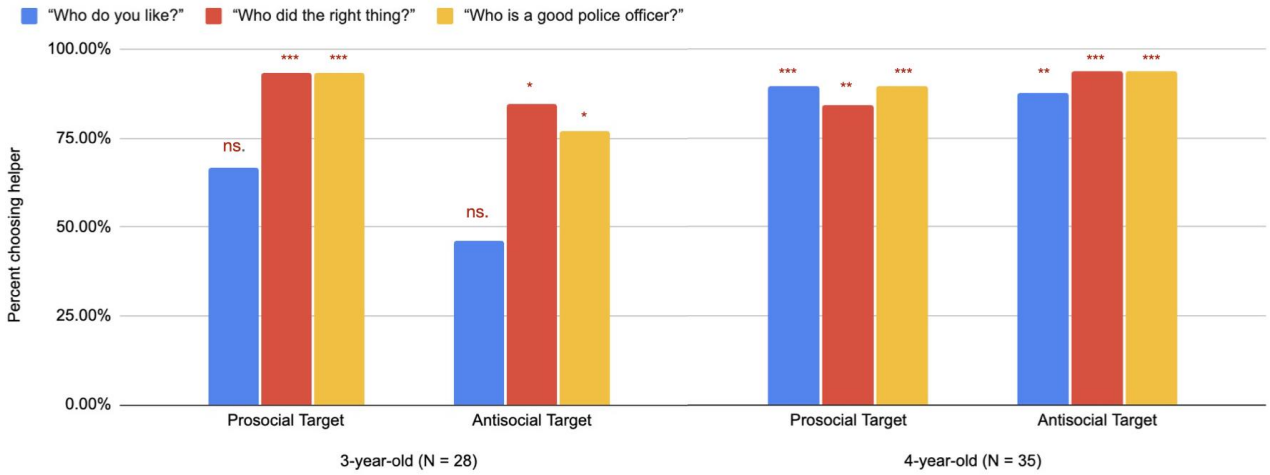

Fig. 2. Percent of 3- and 4-year-old children who chose the rewarder over the hinderer. *$p < .05$, **$p < .01$, *** $p < .001$.

Prosocial vs. antisocial target contexts. One of our main research questions is whether children are sensitive to different contexts in their sociomoral evaluation. To address this question, we examined whether children's sociomoral evaluation for the rewarding and the punishing police officer differed between the prosocial and antisocial target contexts. The scores of the three forced-choice questions were coded as 1 (selecting helper) and 0 (selecting punisher), and then summed for each child (total score ranging between 0 and 3).

We then conducted Mann-Whitney U Test to compare the total scores between the prosocial and antisocial target contexts for each age group. For 3-year-old children, there was no difference in scores between the prosocial target condition and the antisocial target condition, which did not align with our hypothesis; Mann-Whitney U Test, $Z = 1.23$, $p = .22$, $r = .236$. Similarly, for 4-year-old children, we did not detect a significant difference in scores between the prosocial target condition and the antisocial target condition; Mann-Whitney U Test, $Z = .28$, $p = .79$, $r = .046$. Specifically, for children who passed all check questions at the first shot ($N = 36$), the result was the same with the overall sample: there was no difference in scores between the

prosocial target condition and the antisocial target condition, which did not align with our hypothesis; Mann-Whitney U Test, $Z = .53$, $p = .60$, $r = .088$.

Exploratory analyses. To examine whether age, question type, and context influenced children's selection between the rewarding and hindering police officer, we conducted Logistic Regression to analyse the relationship between child's age (3, 4), question type (social evaluation: liking, moral evaluation: rightness and good police officer), and the context (prosocial target, antisocial target), and children's forced-choice selection (children's selections were coded as 1: selecting rewarding police officer; 0: selecting punishing police officer).

The overall logistic regression model was significant, $\chi^2(3) = 13.97$, $p = .003$, Nagelkerke $R^2 = .072$. Question type was found to be significant in predicting the odds of selecting the rewarding police officer. Specifically, the odds of selecting the rewarding police officer would on average increased by 188.6% when asked moral evaluation questions (rightness, good police officer) compared with the social evaluation question (liking), after controlling for age and context, 95% *CI* [0.24, 1.88], $p = .011$. The result also revealed that holding age and question type constant, context was not significant in predicting the odds of selecting the rewarding police officer, 95% *CI* [-2.01, 0.16]), $p = .094$. Holding age and context constant, age was not significant in predicting the odds of selecting the rewarding police officer, 95% *CI* [-0.60, 1.47]), $p = .411$.

Action acceptability ratings. Children's response on the moral scale was measured in order to examine their moral judgment about right and wrong in different contexts. Specifically, we looked at 3- and 4-year-old children's rating of the acceptability of the police officers' rewarding or punishing action with a likert scale ranging from 1 (really bad) to 5 (really good).

To note, given our small sample size and the asymmetrically distributed rating scores in our current sample, we conducted One-sample Wilcoxon Signed-Rank Tests instead of One-sample T-tests to compare children's rating with the baseline (i.e., with score 3: "just ok"). In the antisocial target context, both 3- and 4-year-old children's ratings (3-year-olds: $M = 4.17$, $SD = 1.59$; 4-year-olds: $M = 4.50$, $SD = 1.03$) toward the rewarding police officer's action were significantly more positive than the baseline (i.e., with score 3: "just ok"); One-sample Wilcoxon Signed-Rank Test, 3-year-olds: $Z = 2.70$, $p = .007$, $r = .551$; 4-year-olds: $Z = 4.60$, $p < .001$, $r = .814$. In the antisocial target context, both 3- and 4-year-old children's ratings (3-year-olds: $M = 1.92$, $SD = 1.31$; 4-year-olds: $M = 1.69$, $SD = 1.49$) toward the punishing police officer's action were significantly more negative than the baseline; One-sample Wilcoxon Signed-Rank Test, 3-year-olds: $Z = 2.96$, $p = .003$, $r = .605$; 4-year-olds: $Z = 3.33$, $p < .001$, $r = .589$.

In the prosocial target context, both 3- and 4-year-old children's ratings (3-year-olds: $M = 5.00$, $SD = 0$; 4-year-olds: $M = 4.47$, $SD = 1.12$) toward the rewarding police officer's action were significantly more positive than the baseline; One-sample Wilcoxon Signed-Rank Test, 3-year-olds: $Z = 5.07$, $p < .001$, $r = .976$; 4-year-olds: $Z = 4.56$, $p < .001$, $r = .740$. When rating the punishing police officer in the prosocial target context, both 3- and 4-year-old children's ratings (3-year-olds: $M = 1.64$, $SD = 1.28$; 4-year-olds: $M = 1.21$, $SD = .71$) were significantly more negative than the baseline; One-sample Wilcoxon Signed-Rank Test, 3-year-olds: $Z = 3.51$, $p < .001$, $r = .662$; 4-year-olds: $Z = 5.30$, $p < .001$, $r = .860$.

We also compared 3- and 4-year-old children's rating of rewarding and punishment in both the prosocial and antisocial target context. Due to the non-normal distribution of the data and the small sample size, the Wilcoxon signed-rank test was employed as an alternative to the paired-sample t-test. In the antisocial target context, both 3- and 4-year-olds rated rewarding

(3-year-olds: $M = 4.17$, $SD = 1.59$; 4-year-olds: $M = 4.50$, $SD = 1.03$) more positively than punishment(3-year-olds: $M = 1.92$, $SD = 1.31$; 4-year-olds: $M = 1.69$, $SD = 1.49$); Wilcoxon signed-rank tests: 3-year-olds: $Z = 2.49$, $p = .013$, $r = .787$; 4-year-olds: $Z = 3.18$, $p = .001$, $r = .917$. In the prosocial target context, both 3- and 4-year-olds rated rewarding (3-year-olds: $M = 5.00$, $SD = 0$; 4-year-olds: $M = 4.47$, $SD = 1.12$) as being more positive than punishing (3-year-olds: $M = 1.64$, $SD = 1.28$; 4-year-olds: $M = 1.21$, $SD = 0.71$); Wilcoxon signed-rank tests: 3-year-olds: $Z = 3.23$, $p = .001$, $r = .933$; 4-year-olds: $Z = 3.85$, $p < .001$, $r = .908$.

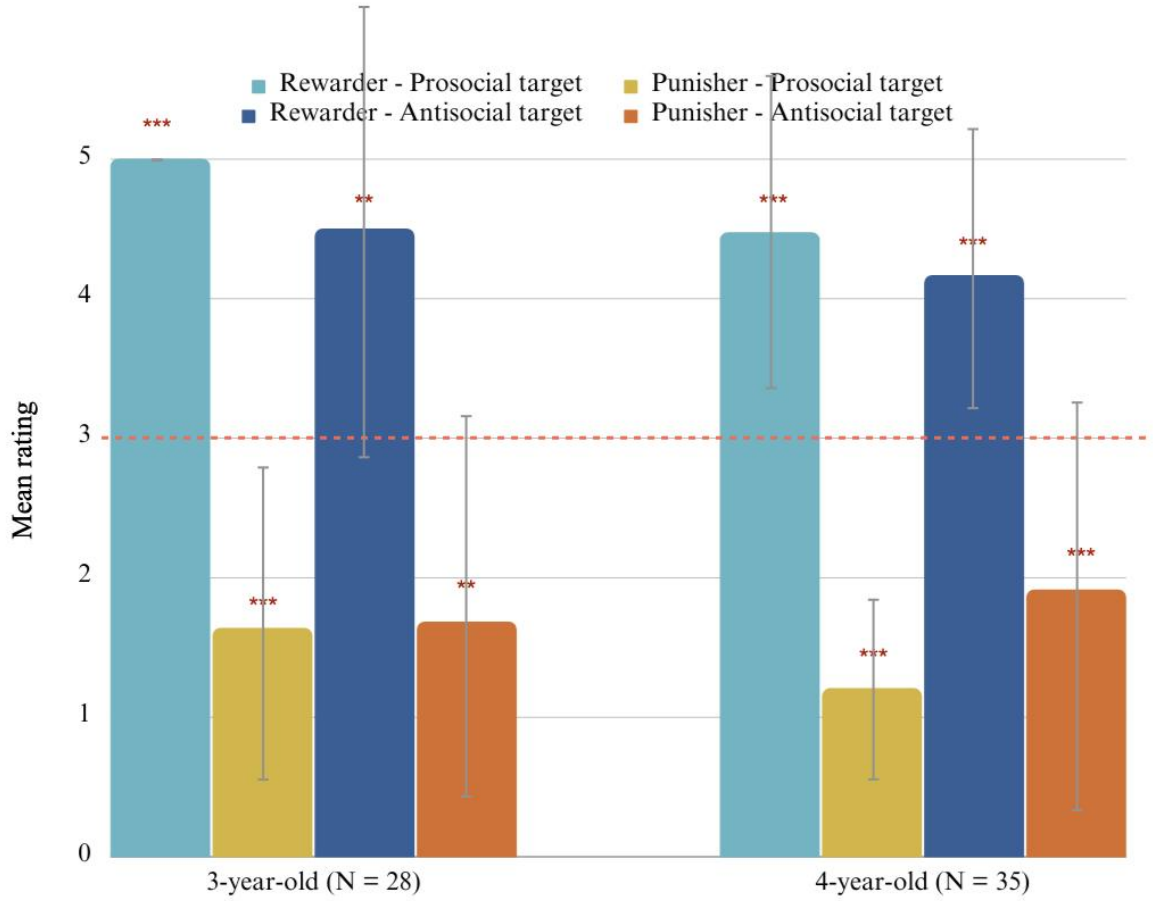

Fig. 3. Mean rating score for the acceptability of the punishing and rewarding action. *$p < .05$, **$p < .01$, *** $p < .001$.

Prosocial vs. antisocial target contexts. We failed to found significant difference of the rating of rewarders between the prosocial target condition and the antisocial target condition for both 3- and 4-year-old children; Mann-Whitney U tests: 3-year-olds: $Z = 1.83$, $p = .067$, $r = .366$; 4-year-olds: $Z = .21$, $p = .835$, $r = .035$. Similarly, both 3- and 4-year-old children's rating of punisher did not differ between the prosocial target condition and the antisocial target condition; Mann-Whitney U tests: 3-year-olds: $Z = .92$, $p = .359$, $r = .180$; 4-year-olds: $Z = .79$, $p = .430$, $r = .133$.

Verbal explanations. We analyzed children's responses regarding the reasons behind their selection for liking, rightness, and a good police officer question (see Table 1). We coded children's answers into four categories: 1) uninformative responses, 2) moral actions and motivations, 3) social actions and motivations, and 4) target's prosocial/antisocial actions.

When asked to explain their selection for liking, 3-year-olds most frequently (53.57%) provided uninformative answers (e.g., "I don't know"); for those who provided informative answers, most of the children focused on the moral values and motivations (42.86%); 4-year-old children most often focused on the police officer's moral actions and motivations (4-year-olds: 62.86%) (e.g., "He gave the ball back." "He is nice."). Similarly, when answering why they thought the selected puppet did the right thing, the majority of children's answers fell into the category of moral actions and motivations (3-year-olds: 67.86%; 4-year-olds: 74.29%). Likewise, when explaining why the selected character was a good police officer, children most often referred to moral actions/motivations (3-year-olds: 64.29%; 4-year-olds: 82.86%).

Table 1

Percentage of children's verbal explanations of each response type

Age	Question	Uninformative	Informative		
			Moral actions and motivations	Social actions and motivations	Target's prosocial/antisocial actions
3	Liking	**53.57%**	42.86%	3.57%	0%
	Rightness	28.57%	**67.85%**	3.57%	0%
	Good police officer	32.14%	**64.29%**	3.57%	0%
4	Liking	8.33%	**91.67%**	0%	0%
	Rightness	20.00%	**74.29%**	5.71%	0%
	Good police officer	14.29%	**85.71%**	0%	0%

Discussion

There has been a debate regarding whether children's sociomoral evaluation is selective and context-dependent; that is, whether children perceive helping as always good and hindering as always bad. Evidence showed that adults tend to be selective in contexts and positively evaluate those who punish antisocial others (Barclay, 2006; Gordon, Madden, & Lea, 2014; Gintis, Smith, & Bowles, 2001; Nelissen, 2008; Martin et al., 2019). Although some evidence suggested children would consider the target's previous prosocial or antisocial actions in their socio-moral evaluation (Geraci, 2021; Hamlin et al., 2011; Lee & Warneken, 2020; Loke et al., 2011), some showed they always preferred helpers and evaluated helping as good and hindering as bad (Li et al., 2020; Li & Tomasello, 2018; Van de Vondervoort, 2020). The aim of the current study was to investigate whether preschool-age children are sensitive to contextual information when evaluating third-party intervention when the rewarder and the punisher held authority status. In this study, 3- and 4-year-old children watched live puppet shows featuring both prosocial and antisocial events. They then observed a rewarding authority figure providing help to either the previously prosocial or antisocial target, while a punishing authority figure hindered the same target. After the show, we asked the children to choose between the rewarding and punishing authority figures based on their evaluations of liking, rightness, and the good police officer. We also assessed the children's ratings of the appropriateness of the authority figures' helping and punishing actions using a five-point likert scale.

The results showed that the children's sociomoral evaluations of authority figures' rewarding and punishing actions were not influenced by the context: regardless of whether the target had previously acted prosocially or antisocially, children always preferred to select the

helper police officer over the punishing police officer regarding liking, rightness, and who was the good police officer. Moreover, children evaluated the rewarding police officer positively and the punishing police officer negatively in both the prosocial and antisocial target context, despite they themselves thought the antisocial target should be punished (i.e., selecting the hinderer over the helper when asked "who should get in trouble?"). These results aligned with previous work which revealed preschool-age children's judgement of third-party intervention was not influenced by contextual information (Li & Tomasello, 2018; Van de Vondervoort, 2020). When asked to justify their selections in terms of social preference (i.e., liking) and moral evaluation (i.e., rightness, good police officer), aligning with the result of a similar study (Van de Vondervoort, 2020), the majority of both 3- and 4-year-old children referred to the moral values and motivations of the helping or punishing police officers' actions, which might suggest that 3- and 4-year-olds interpreted the puppet shows and the characters' actions in terms of their moral values.

To note, in the current study, children were asked comprehension questions and successfully passed memory checks about the target's previous action after the box shows and before each ball shows. Therefore, we could exclude the possibility that children's failure to consider the targets' previous actions were due to their misunderstanding of the show, insufficient attention or memory loss. In addition, children who passed all the check questions on the first try showed the same pattern of choice and evaluation as the whole sample. This suggested that confounding factors, such as differences in levels of engagement with the puppet show and memory capacity, did not skew the results. In conclusion, the result suggested that 3- to 4-year-old children might be insensitive to contextual information when evaluating third-party punishment, even when the punishers held authority status.

Our result showed that in both prosocial and antisocial target contexts, 3-year-old children's selection of the rewarding police officer did not significantly differ from chance-level in terms of their social preference ("liking") , but their selection of the rewarding police officer for moral preferences ("rightness", "good police officer") were significantly above chance-level. However, 4-year-old children consistently preferred the rewarding police officer across all three questions in both prosocial and antisocial target contexts. It was also found that both 3- and 4-year-old children were more likely to select the rewarding police officer when asked moral evaluation questions (rightness, good police officer) than when asked the social evaluation question (liking).

One explanation is that children interpret the current puppet show as being more focused on moral aspects rather than social aspects. This interpretation is supported by the fact that the majority of both 3- and 4-year-old children's verbal responses, for both the social preference question and the puppet show, were centered around the police officers' moral actions and motivations. This suggests that the children may have perceived the characters' behaviors through a moral lens rather than considering their implications for social preferences. Moreover, compared to the 4-year-olds, 3-year-old children might be more cognitively challenged in generalizing social-related judgments based on the police officers' moral actions. This could hinder their ability to incorporate the police officers' actions when generating their social preferences, leading to their confusion when asked about their social preference. This explanation could be supported by children's verbal responses: when asked about their social preference (liking), a majority of 3-year-olds provided uninformative responses (53.57%), suggesting that children might be confused about the social preference question.

Why did children disapprove of punishment enforced by authority figures?

In the current study, children exhibited a preference for those who rewarded wrongdoers over those who punished wrongdoers, and did not judge punishment toward antisocial individuals as less appropriate than rewarding, even when the punisher was depicted as an authority figure who should be in a position to punish. Combining with the findings in Van de Vondervoort's (2020) study, our results suggest that children would more negatively evaluate third-party punishment than rewarding regardless of the agent's authority status. Our result contrasts with relevant studies which suggest children and adults would expect authority figures to punish and judge punishment from an authority figure as appropriate. For example, compared to ordinary citizens, adults judge authority figures as more obligated to engage in third-party punishment, and being more deserving of punishment when they did not engage in punishment when they should have (Martin et al., 2019). A developmental study revealed that 17-month-old infants expect leaders to intervene in transgressions and rectify wrongdoing, but they have no such expectation for a non-leader individual (Stavans & Baillargeon, 2019). Additionally, 4- to 7-year-old children judged authority figures as not only obligated to intervene, but also obligated to punish when witnessing transgressions (Marshall, Mermin-Bunnell, & Bloom, 2020). To sum up, the result of the current study contrasted with many existing studies that supported the idea that children and adults are sensitive to the authority status of punishers, expecting and endorsing authority figures to take on the responsibility of enforcing moral norms, while holding no such expectation for non-authority figures.

What might have caused the discrepancy between our result and the existing evidence in evaluating authority figures' third-party punishment? One may argue that it was because 3- to

4-year-old children have not yet developed a conceptual understanding of authority figures or of social hierarchies, or the authority status was not made salient in our puppet show. However, in our study, the character's authority status was repeatedly stressed, and children demonstrated a clear understanding of the authority figure and the obligation associated with their social identity, which is to punish transgressors. For instance, the helper and punisher were introduced and repeatedly referred to as police officers, wearing police uniforms. When asked about what police officers do, many children correctly pointed out the obligation to enforce norms of police officers (e.g., "They catch bad guys!"). For children who failed to clearly identify the police officers' responsibility in their verbal response, we emphasized the norm-enforcing obligation associated with the police officer role before we proceeded to the puppet shows ("Police officers make sure everyone follows the rules!"). Therefore, we can eliminate the possibility that children lack a conceptual understanding of authority figures and the punitive obligation associated with their authority status.

However, considering that children in our sample were mostly recruited from high socioeconomic status (SES) families, their experience of the sanctioning responsibility of police officers was extremely limited, mostly from hearsay or from stories and cartoons, rather than real-life, first-hand experience. Moreover, empirical evidence found that people of high SES tend to hold a more collaborative and less coercive perspective of power compared with people from low SES background (Ten Brinke & Keltner, 2022). It is possible that the punitive obligations of police officers were not salient in their understanding of the role of police officers. As evidence, many of the children in our sample responded with "They help people!" when asked about the responsibilities of police officers. Therefore, it is possible that children in our sample did not consider the punitive responsibility as crucial, even though they did indeed understand it, in their

evaluation of the actions of the punishing police officers in our puppet show. Future studies can examine other types of authority figures, such as teachers and parents, with whom children have more real-life experience and a deeper understanding of the disciplinary responsibilities associated with these authority roles. Another interesting direction is to compare children with a high SES background recruited from areas with low crime rates, to children with a low SES background living in neighborhoods with high crime rates. By doing so, we can examine whether the real-life experience of the punitive practices of police officers would influence their perception of the punitive obligation associated with the role of police officers, and their sociomoral evaluation of third-party punishment of antisocial individuals.

In addition, it is possible that the different results between the past studies and the current study could be due to the different ways of measuring the evaluation of authority figures' third-party punishment. For instance, in Marshall, Mermin-Bunnell, & Bloom's (2020) study, they presented 4- to 7-year-olds with a story about a transgression and measured their evaluation by asking whether the authority figure should get the transgressor in trouble (e.g., "Do you think Emma should get Jessica in trouble for saying something mean?"). Thus, they measured children's evaluation of an imaginary punishment. In contrast, our study presented children with both the transgression and the punishing action and result, which may have led children to focus more on the nature of the punishment and its outcome (since it is moral salient), rather than the authority status of the punishing agent. Thus this difference in measurement approaches might explain the conflicting findings between our study and the existing evidence. It is worth noting that the majority of 3- and 4-year-olds emphasized moral values and motivations in their verbal responses regarding their choice of liking, perception of rightness, and who is a good police officer. In contrast, none of the children in our sample referred to the authority status of the

police officer or the police officer's obligation to punish when explaining their evaluations. We can infer that children in the current study might have focused mainly on the intervention method, and did not take the agent's authority status into consideration when judging the appropriateness of the punishment.

Why did children appear non-selective in their moral evaluation?

Another key question to discuss based on our results is why preschoolers disapprove of others' third-party punishment when they themselves deem the transgressor as deserving of punishment. In the current study, while children themselves believed the antisocial target should be punished (i.e., uniformly selecting the hinderer over the helper when asked "who should get in trouble?"), they still preferred and positively evaluated the rewarding police officers and negatively evaluated the punishing police officers in the antisocial target context.

Although preschool-age children voluntarily enforce punishment as third parties (McAuliffe, Jordan, & Warneken, 2015; Kenward & Dahl, 2011; Kenward & Östh, 2015; Rossano, Rakoczy, & Tomasello, 2011; Yudkin, Van Bavel, & Rhodes, 2020), evidence regarding whether children would positively or negatively evaluate third-party punishers based on context was mixed. On one hand, some studies found that preschool-age children always positively evaluate those who help and negatively evaluate those who punish, without considering whether the action is directed to prosocial or antisocial others (Li & Tomasello, 2018; Van de Vondervoort, 2020), which align with our result. On the other hand, our result contrasted with many evidence that children would expect third parties (especially authorities) to punish wrongdoers, preferred third parties who enforced punishment on transgressors, and evaluated punishers more positively (Geraci, 2021; Hamlin et al., 2011; Lee & Warneken, 2020;

Loke et al., 2011; Marshall, Mermin-Bunnell, & Bloom, 2020; Vaish et al., 2016). What factors could explain children's failure to consider contextual information in our study?

Explanation 1: Fail to capture the developmental transition

One possible explanation for the current result is that, at 3 to 4 years of age, children indeed have not developed the capacity to integrate the contextual information, or to consider contextual information as crucial when making evaluations, and thus were aversive to punishment regardless of the context.

It is possible that only older children, but not younger children, would be able to integrate the contextual information of the helping/hindering action in their sociomoral evaluations. For example, the result in our study aligned with a recent study, where they found that 2- to 4-year-old children uniformly evaluated helping as positive and hindering as negative without considering the nice or mean goal of the recipient. However, in the same study, children aged 5- to 7-year-old would consider the moral nature of the action's goal, preferring agents who hindered immoral actions over those who helped (Myslińska Szarek, Baryła, & Wojciszke, 2023). Similarly, another piece of evidence demonstrated that 5-year-olds, but not 4-year-olds, judged enforcers and enforcement more positively than non-enforcers and non-enforcement, personally preferred enforcers to non-enforcers, and provided more material rewards to enforcers than non-enforcers (Vaish et al., 2016). Studies on older children found that children through 5- to 9-years of age positively evaluated third-party punishers who sanctioned unfair allocators (Lee & Warneken, 2020). Such evidence indicated that the capacity to appreciate altruistic punishment and those who punish may develop between the ages of 4 and 5. This transitional age in development is consistent with developmental evidence on the development of relevant

capacities such as group norms and group loyalty. For instance, it was discovered that 5-year-olds exhibit a strong understanding of complex cooperation and group norms, whereas 4-year-olds do not (Misch, Over, & Carpenter, 2014).

There could be two explanations for the different capability of analysing contextual information between younger and older children. Firstly, at 3 and 4 years old, children have just started attending preschool and are in the process of learning various social norms from their teachers, such as the idea that "two wrongs don't make a right." As a result, they may be particularly sensitive to all forms of antisocial behavior, even though certain negative actions, such as altruistic punishment, serve a prosocial purpose. For example, a study examined whether 4-year-old children would identify more with (i.e., choosing to re-enact the role of) a punisher of antisocial others or a non-punisher. The results revealed that despite that children agreed that the transgressors should be punished, they did not identify more with punishers who sanctioned wrongdoers than non-punishers (Kenward & Östh, 2012). We may infer from such evidence that preschool-age children might not appreciate the prosocial purpose of punishment and value punishers. Secondly, younger children's failure to consider contextual information may derive from their limited cognitive capacities. Past studies revealed that preschool-age children were not cognitively able to make moral evaluations based on the agent's motive when the motive and outcome were incongruent in valence, while 2nd graders were able to (Nelson, 1980). Similarly in our study, even though 3- and 4-year-old children correctly identified the antisocial target as bad, they might not be able to connect and integrate both the police officers' intervention and the target's previous action while making sociomoral evaluations.

Based on this explanation, we may infer that one of the crucial limitations of the current study is that the age range included in our sample is rather small. Specifically, we only tested 3-

and 4-year-olds. This limitation might have caused the results' failure to detect a clear developmental trajectory of children's moral evaluation of authority figures' third-party punishment, assuming the transition in development happens at between age 4 and 5. Future studies could recruit 5-year-old children and compare their evaluations of authority figures' third-party punishment with those of the 3- to 4-year-old children in our sample. This would help provide a better understanding of the developmental trajectory in children's moral evaluations.

Explanation 2: Not the specific punishing method

It is also possible that children in the current study are indeed capable of considering the context when evaluating punishment, but they appeared to be insensitive to contextual information in their sociomoral evaluation because they perceived the specific punishment method depicted in the current puppet show (i.e., taking someone's belongings away) as unacceptable or inappropriate.

Generally, third-party interventions related to punishment can take various forms, including direct punishment (e.g., sending someone to "time-out"), indirect punishment (e.g., gossiping), recruiting others to intervene (e.g., reporting to a teacher), and protesting. Moreover, third parties are not limited to intervening through punishment. There are alternative forms of non-punitive third-party intervention, such as rectifying wrongs (e.g., compensating the victim), partner-choice (e.g., ending the relationship with the wrongdoer), and encouraging forgiveness (Marshall & McAuliffe, 2022). In the current study, the punishing police officer hindered the target through direct punishment, namely taking away the target's ball. This form of punishment may be perceived as excessively harsh and unacceptable in children's perspective. It is possible that children might still be sensitive to contextual information and positively evaluate certain

forms of punishment when the target was antisocial. And our failure to detect such capability might be because children were sensitive to the method of punishment, and may deem the specific way of punishment in the current puppet show as inappropriate.

Therefore, understanding the nuances of children's acceptability of different punitive methods is crucial. Among the current empirical evidence that children endorse third-parties who punish transgressors, the method of intervention is milder compared to our approach, such as punishing through verbally accusing (Vaish at al., 2016), reporting the transgression to authority figures (Loke et al., 2011; Loke et al., 2014), and intervening through protecting victims (Kanakogi et al., 2017). When asked to rate four parental disciplinary methods (i.e., spanking, reasoning, withdrawing privileges, and time-out), 6- to 10-year-old children rated reasoning as the most fair, and corporal punishment through spanking as least fair and not for long-term behavior change (Vittrup & Holden, 2010).

Evidence suggested that children tend to evaluate mild punishment as more positive compared to hard punishment, and would consider the severity of the transgression when evaluating the appropriateness of punishment. For instance, 4- to 6-year-old children evaluated punishers who enforced milder punishment (i.e., removing 1 or 2 objects) more positively than those who imposed harsher punishment (i.e., removing 3 objects) (Liu, Yang, & Wu, 2021). In Loke et al. 's (2011) study, 8- to 11-year-old children positively evaluated reporting major transgressions to teachers but negatively evaluated reporting minor transgressions to teachers. In contrast, 6- to 7-year-olds were insensitive to the seriousness of the transgression, and positively evaluated those who reported transgressions to teachers in both major and minor transgression scenarios. Such results suggest that as children grow older, they may become more sensitive to the seriousness of transgressions when judging whether a punishment is appropriate.

It is also found that at preschool age, children prefer other forms of third-party intervention than punishment. While punishment is effective in discouraging transgressions and sustaining cooperation, restoration could also help resolving conflicts and minimizing harm. More importantly, even without considering its global purpose, restoration itself is a prosocial action while punishment itself is a negative action. Evidence suggested preschool-age children might prefer restorative justice over punishing the antisocial individual. For example, in one study, 3-year-old children were given the opportunity to remove items and prevent the wrongdoer from gaining a reward, as well as the opportunity to restore the lost items for the victim. The result revealed that children preferred to choose restoration over removal as third-party intervention (Riedl et al., 2015). 4- to 6-year-olds evaluated individuals who restored the belongings more positively than those who punished the transgressor (Liu, Yang, & Wu, 2021). In another study, although children positively evaluated both third-party helpers who compensated the victim and third-party punishers who took resources away from the transgressor, they preferred helpers over punishers, and preferred helping over punishment when asked to assess the type of third-party intervention (Lee & Warneken, 2020). To sum up, these findings suggested that preschool-age children might be motivated by reducing inequality, rather than seeking retribution, when evaluating third-party interventions in response to transgressions. To better understand whether children are sensitive to contextual information in third-party intervention, future studies could explore children's evaluations of third-party's restorative behavior as an alternative to punishment.

Together, preschool-age children are sensitive to the method and severity of punishment, and they may only deem certain forms of third-party punishment as appropriate. Therefore, the current study is limited in that we only explored one punitive approach (i.e., taking away the

target's belongings). Children tested in the current study might perceive this punitive method as too harsh and antisocial. Future studies could explore other more lenient forms of punishment, such as sending the target for a time-out, preventing the target from joining the game, or verbally scolding the wrongdoer, in order to explore children's context-dependent evaluations of punishment.

Explanation 3: Dislike the bad consequences of punishment

Furthermore, it is possible that preschoolers were indeed selective in judging the deservingness of punishment, but they appeared non-selective in judging the appropriateness of punishment in the current study because they disliked the negative consequences of punishment. In Li & Tomasello's (2018) study, children aged 3 and 5 watched a video-recorded puppet show where a prosocial individual shared food and an antisocial individual hit their social partners. Subsequently, the children observed four intervention events in which an agent attempted to help or hinder the previously prosocial or antisocial puppet in opening a box, with successful or failed intervening results. The results revealed that while 3-year-olds negatively evaluated those who successfully punished the antisocial target, they positively rated punishers who intended but failed to punish. The same results were replicated in another study involving 3-year-old German children using the same experimental procedure (Li et al., 2020), suggesting that generally 3-year-olds may believe that third-parties should signal they care when witnessing a transgression, but they dislike the bad consequence when transgression is enforced.

Such evidence suggested that preschool age children may be sensitive to contextual information, recognizing that norm violations should be punished and appreciating individuals who demonstrate the intention to sanction antisocial actions. However, considering the fact that

in the current study, both the helping and punishing scenarios were successful interventions, they may simply dislike the negative consequences of punishment, and thus negatively evaluate successful punishers. This possibility aligns with evidence that children are selective in their punitive motivations, as reflected in their evaluation of the deservingness of punishment, as well as their own punitive actions (McAuliffe, Jordan, & Warneken, 2015; Kenward & Dahl, 2011; Kenward & Östh, 2015; Rossano, Rakoczy, & Tomasello, 2011; Yudkin, Van Bavel, & Rhodes, 2020), but they seem to be non-selective in evaluating successful punishment enforced by others (Li & Tomasello, 2018; Li et al., 2020; Van de Vondervoort, 2020). Future studies could examine how different consequences affect children's social and moral evaluation of third-party interventions, and explore how the intention and outcome of punishment interact in children's evaluations of authority's third-party punishment.

Explanation 4: Small sample size

Additionally, one obvious limitation of the current study is that the sample size is rather small. With only twenty-nine 3-year-olds and thirty-five 4-year-olds in our sample, and utilizing a 2 (age; between person) × 2 (prosocial versus antisocial target; between person) × 2 (punish versus reward; within person) mixed design, we ended up having only thirteen to nineteen children in each of the four subgroups. The effect size of the current data analysis is small to medium, which suggests that with our small sample size, our sample may not have sufficient power to detect the effect of interest, considering the possible sampling error associated with small sample size.

To summarize, the current study aimed to clarify whether preschool-age children would consider contextual information and judge altruistic punishment as appropriate when the

punisher held authority status. Specifically, we portrayed the punisher as a police officer, in order to rule out the possibility that children's disapproval of altruistic punishment stemmed from a belief that ordinary individuals lacked the right to punish. Our findings showed that children appeared insensitive to contextual information, always preferred rewarding police officers over punishing police officers and uniformly evaluated punishment as negative and rewarding as positive, even when the punisher held an authoritative position and were identified as obligated to punish by our participants. Our study shed light on how authority status influences children's sociomoral evaluation toward third-party punishment, and deepened our understanding of whether children are able to integrate contextual information in sociomoral evaluation. Additionally, we discussed possible explanations of our results. These included children's perception of the punitive responsibility of authority figures in our sample, the possible age-related difference between young children and old children, the inappropriate method of punishment, the influence of our small sample size, and the notion that children may be able to selectively form an intention to punish based on the target's previous action, while being reluctant to witness negative outcomes. We also outlined future directions to better explore the contexts in which children might appreciate third-party punishment. The findings contribute to our understanding of how contextual information can influence children's evaluations of third-party punishment and punishers, and provide insights into the developmental trajectory of children's moral sense.

References

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

Behne, T., Carpenter, M., Call, J., & Tomasello, M. (2005). Unwilling versus unable: Infants' understanding of intentional action. *Developmental Psychology*, 41, 328–337.

Bendor, J., & Swistak, P. (2001). The evolution of norms. *American Journal of Sociology*, 106(6), 1493-1545. doi: 10.1086/321298

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. https://doi.org/10.1073/pnas.0630443100

Brown, D. E. (1991). *Human universals.* Temple University Press.

Bull, J. J., & Rice, W. R. (1991). Distinguishing mechanisms for the evolution of cooperation. *Journal of Theoretical Biology*, 149(1), 63-74. doi: 10.1016/s0022-5193(05)80072-4.

Clutton-Brock, T., Parker, G. (1995). Punishment in animal societies. *Nature* 373, 209–216 . https://doi.org/10.1038/373209a0

Eriksson, K., Andersson, P. A., & Strimling, P. (2016). Moderators of the disapproval of peer punishment. *Group Processes & Intergroup Relations*, 19(2), 152–168. https://doi.org/10.1177/1368430215583519

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

Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature*, 425(6960), 785-791. https://doi: 10.1038/nature02043.

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, 25(2), 63-87. https://doi.org/10.1016/S1090-5138(04)00005-4

Fessler, D. M., & Haley, K. J. (2003). The strategy of affect: Emotions in human cooperation 12. *The Genetic and Cultural Evolution of Cooperation*, P. Hammerstein, ed, 7-36.

Gardner, A., & West, S. A. (2004). Cooperation and punishment, especially in humans. *The American Naturalist*, 164(6), 753-764. doi: 10.1086/425623.

Geraci, A. (2021). Toddlers' expectations of corporal third-party punishments against the non-defender puppet. *Journal of Experimental Child Psychology*, 210, 105199. https://doi.org/10.1016/j.jecp.2021.105199

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

Gintis, H., Smith, E. A., & Bowles, S. (2001). Costly signaling and cooperation. *Journal of Theoretical Biology*, 213(1), 103-119. https://doi.org/10.1006/jtbi.2001.2406

Hamilton, W. D. (1964). Genetical evolution of social behavior I and II. *Journal of Theoretical Biology.* 7, 1–52. https://doi.org/10.1016/0022-5193(64)90038-4

Hamlin, J. K. (2013). Moral Judgment and Action in Preverbal Infants and Toddlers: Evidence for an Innate Moral Core. *Current Directions in Psychological Science*, 22(3), 186–193. https://doi.org/10.1177/0963721412470687

Hamlin, J. K. (2015). The case for social evaluation in preverbal infants: Gazing toward one's goal drives infants' preferences for Helpers over Hinderers in the hill paradigm. *Frontiers in Psychology*, 5. https://doi.org/10.3389/fpsyg.2014.01563

Hamlin J. K., & Wynn K. (2011). Young infants prefer prosocial to antisocial others. *Cognitive Development*, 26, 30–39. doi: 10.1016/j.cogdev.2010.09.001

Hamlin, J. K., Wynn, K., Bloom, P. (2007). Social evaluation by preverbal infants. *Nature*, 450, 557–559. https://doi.org/10.1038/nature06288

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. https://doi: 10.1111/j.1467-7687.2010.00951.x

Hamlin, J. K., Wynn, K., Bloom, P., & Mahajan, N. (2011). How infants and toddlers reach to antisocial others. *Proceedings of the National Academy of Sciences*, USA, 108, 19931–19936. https://doi.org/10.1073/pnas.1110306108

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

Henrich, J., et al. (2006). Costly punishment across human societies. *Science*, 312(5781), 1767-1770. https://doi: 10.1126/science.1127333.

Henrich, J., et al. (2010). Markets, religion, community size, and the evolution of fairness and punishment. *Science*, 327(5972), 1480–1484. https://doi.org/10.1126/science.1182238

Hilbe, C., Schmid, L., Tkadlec, J., & Nowak, M. A. (2018). Indirect reciprocity with private, noisy, and incomplete information. *Proceedings of the National Academy of Sciences*, 115(48), 12241-12246. https://doi.org/10.1073/pnas.1810565115

Kanakogi, Y., Inoue, Y., Matsuda, G., Itakura, S., & Fukui, Y. (2017). Preverbal infants affirm third-party interventions that protect victims from aggressors. *Nature Human Behaviour*, 1, 0037. https://doi.org/10.1038/s41562-016-0037

Kenward, B., & Dahl, M. (2011). Preschoolers distribute scarce resources according to the moral valence of recipients' previous actions. *Developmental Psychology*, 47(4), 1054–1064. https://doi.org/10.1037/a0023869

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

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

Lee, Y. E., & Warneken, F. (2020). Children's evaluations of third-party responses to unfairness: Children prefer helping over punishment. *Cognition*, 205, 104374. https://doi.org/10.1016/j.cognition.2020.104374

Li, J., Hou, W., Zhu, L., & Tomasello, M. (2020). The development of intent-based moral judgment and moral behavior in the context of indirect reciprocity: A cross-cultural study. *International Journal of Behavioral Development*, 44(6), 525–533. https://doi.org/10.1177/0165025420935636

Li, J., & Tomasello, M. (2018). The development of intention-based sociomoral judgment and distribution behavior from a third-party stance. *Journal of Experimental Child Psychology*, 167, 78-92. https://doi.org/10.1016/j.jecp.2017.09.021

Liu, X., Yang, X., & Wu, Z. (2021). To punish or to restore: How children evaluate victims' responses to immorality. *Frontiers in Psychology*, 12, 696160. https://doi.org/10.3389/fpsyg.2021.696160

Loke, I. C., Heyman, G. D., Forgie, J., McCarthy, A., & Lee, K. (2011). Children's moral evaluations of reporting the transgressions of peers: Age differences in evaluations of tattling. *Developmental Psychology*, 47(6), 1757-1762. doi: 10.1037/a0025357.

Loke, I., Heyman, G. D., Itakura, S., Toriyama, R., & Lee, K. (2014). Japanese and American children's moral evaluations of reporting on transgressions. *Developmental Psychology*, 50, 1520-1531.

Marshall, J., McAuliffe, K. (2022). Children as assessors and agents of third-party punishment. *Nature Reviews Psychology*, 1, 334–344. https://doi.org/10.1038/s44159-022-00046-y

Marshall, J., Mermin-Bunnell, K., & Bloom, P. (2020). Developing judgments about peers' obligation to intervene. *Cognition*, 201, 104215. https://doi.org/10.1016/j.cognition.2020.104215

Martin, J. W., Jordan, J. J., Rand, D. G., & Cushman, F. (2019). When do we punish people who don't? *Cognition*, 193, Article 104040. https://doi.org/10.1016/j.cognition.2019.104040

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

Meristo, M., & Surian, L. (2013). Do infants detect indirect reciprocity? *Cognition*, 129(1), 102-113. https://doi.org/10.1016/j.cognition.2013.06.006

Misch, A., Over, H., & Carpenter, M. (2014). Stick with your group: Young children's attitudes about group loyalty. *Journal of Experimental Child Psychology*, pp. 19-36. ISSN 0022-0965. https://doi.org/10.1016/j.jecp.2014.02.008

Myslińska Szarek, M. K., Baryla, W., & Wojciszke, B. (2023). Is helping always morally good? Study with toddlers and preschool children. *Developmental psychology*, 59(5), 918–927. https://doi.org/10.1037/dev0001521

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

Nelson, S. A. (1980). Factors influencing young children's use of motives and outcomes as moral criteria. *Child Development*, 51(3), 823–829. https://doi.org/10.2307/1129470

Nowak, M. A., & Sigmund, K. (1998). The dynamics of indirect reciprocity. *Journal of theoretical biology,* 194(4), 561–574. https://doi.org/10.1006/jtbi.1998.0775

Rai, T., & Fiske, A. P. (2011). Moral psychology as regulating relationships: Moral motives for unity, hierarchy, equality, and proportionality in social-relational cognition. *Psychological Review*, 118(1), 57–75.

Raihani, N. J., & McAuliffe, K. (2012). Human punishment is motivated by inequity aversion, not a desire for reciprocity. *Biology Letters*, 8(5), 802-804. DOI: 10.1098/rsbl.2012.0470

Riedl, K., Jensen, K., Call, J., & Tomasello, M. (2015). Restorative Justice in Children. *Current biology* : CB, 25(13), 1731–1735. https://doi.org/10.1016/j.cub.2015.05.014

Rossano, F., Rakoczy, H., & Tomasello, M. (2011). Young children's understanding of violations of property rights. *Cognition*, 121(2), 219–227. https://doi.org/10.1016/j.cognition.2011.06.007

Scola, C., Holvoet, C., Arciszewski, T., & Picard, D. (2015). Further evidence for infants' preference for prosocial over antisocial behaviors. *Infancy*, 20(6), 684–692. https://doi.org/10.1111/infa.12095

Sober, E., & Wilson, D. S. (1998). Unto others: The evolution and psychology of unselfish behavior. Harvard University Press.

Stavans, M., & Baillargeon, R. (2019). Infants expect leaders to right wrongs. *Proceedings of the National Academy of Sciences of the United States of America*, 116(41), 16292–16301.https://doi.org/10.1073/pnas.1820091116

Ten Brinke, L., & Keltner, D. (2022). Theories of power: Perceived strategies for gaining and maintaining power. *Journal of personality and social psychology*, 122(1), 53–72. https://doi.org/10.1037/pspi0000345

Thaler, R. H. (1988). Anomalies: The ultimatum game. *Journal of Economic Perspectives*, 2, 195-206.

Vaish, A., Carpenter, M., & Tomasello, M. (2009). Sympathy through affective perspective taking and its relation to prosocial behavior in toddlers. *Developmental Psychology*, 45(2), 534-543. https://doi.org/10.1037/a0014322

Vaish, A., Herrmann, E., Markmann, C., & Tomasello, M. (2016). Preschoolers value those who sanction non-cooperators. *Cognition*, 153, 54-58. doi:10.1016/j.cognition.2016.04.011

Vaish, A., Missana, M., & Tomasello, M. (2011). Three-year-old children intervene in third party moral transgressions. *The British Journal of Developmental Psychology*, 29, 124–130. doi:10.1348/026151010X532888

Van de Vondervoort, J. W. (2020). Young children's social and moral evaluations of third-party helpers and hinderers. University of British Columbia. doi: 10.14288/1.0388328

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 de Vondervoort, J. W., & Hamlin, J. K. (2018). Preschoolers focus on others' intentions when forming sociomoral judgments. *Frontiers in Psychology*, 9, 1851.

Vittrup, B., & Holden, G. W. (2010). Children's assessments of corporal punishment and other disciplinary practices: The role of age, race, SES, and exposure to spanking. *Journal of Applied Developmental Psychology*, 31(3), 211-220.

Wedekind, C., & Milinski, M. (2000). Cooperation through image scoring in humans. *Science*, 288(5467), 850-852. DOI: 10.1126/science.288.5467.850

Yamagishi, T. (1986). The Provision of a Sanctioning System as a Public Good. *Journal of Personality and Social Psychology*, 51(1), 110-116. https://doi:10.1037/0022-3514.51.1.110

Yudkin, D. A., Van Bavel, J. J., & Rhodes, M. (2020). Young children police group members at personal cost. *Journal of Experimental Psychology: General*, 149(1), 182–191. https://doi.org/10.1037/xge0000613