

# Learning Loopholes: The Development of Intentional Misunderstandings in Children

Sophie Bridgers<sup>1,2\*</sup>, Kiera Parece<sup>1\*</sup>, Ibuki Iwasaki<sup>1</sup>, Annalissa Broski<sup>1</sup>, Laura Schulz<sup>1</sup>, and Tomer Ullman<sup>2</sup>

<sup>1</sup>Department of Brain and Cognitive Sciences, MIT

<sup>2</sup>Department of Psychology, Harvard University

\*These authors contributed equally to this work.

## Abstract

What do children do when they do not want to obey but cannot afford to disobey? Might they, like adults, feign misunderstanding and seek out loopholes? Across four studies (N=733; 45% female; USA; majority White; data collected 2020-2023), we find that loophole behavior emerges around ages 5 to 6 (Study 1, 3-18 yrs), that children think loopholes will get them into less trouble than non-compliance (Study 2, 4-10 yrs), predict that other children will be more likely to exploit loopholes when goals conflict (Study 3, 5-10 yrs), and are increasingly able to generate loopholes themselves (Study 4, 5-10 yrs). This work provides new insights on how children navigate the gray area between compliance and defiance and the development of loophole behavior across early and middle childhood.

Children are often in a bind: They don't always want to obey adults, but they can rarely afford not to. What is a self-respecting child to do? Consider a parent who tells his daughter happily playing on the floor amid a sea of toys: "When I come back I don't want to see anything on the floor." A goody-two-shoes would clean everything up and put

it away where it belongs. An ungovernable rebel would ignore the request and keep right on playing. But a resourceful trickster might ensure that, when the parent comes back, everything is indeed off the floor – and the child is happily playing amid the sea of toys now on top of her bed. The parent’s request has been met. Technically.

When simply obeying is unappealing, but outright disobeying is risky, the ambiguity of language can provide an out in the form of a ‘loophole’ (Bridgers et al., 2023). A loophole identifies an alternative interpretation of a request that better aligns with one’s own goals, and can be a means to still do what one wants (e.g., continue playing with toys), while avoiding the negative consequences of direct non-compliance (e.g., missing out on dessert). Recent research with adults reveals that they exploit loopholes in situations of goal conflict, especially when their social partner is of equal or higher status than they are (Bridgers et al., 2023). Adults also predict loopholes will be less upsetting than non-compliance, and less likely to get people into trouble (Bridgers et al., 2023; Qian, Bridgers, Taliaferro, Parece, & Ullman, 2024). Even in the court of law, legal loopholes that follow the letter but not the spirit of a law can be exonerating, or at least lead to more lenient sentencing (Garcia, Chen, & Gordon, 2014; Hannikainen et al., 2022; Isenbergh, 1982; Katz, 2010; Struchiner, Hannikainen, & de Almeida, 2020). Beyond its function in social interactions, loophole behavior is also a cultural object. Loopholes are so often woven into centuries-old fables of people outwitting malevolent forces through clever misinterpretations, or being similarly tricked by a mischievous spirit, that such stories form a genre of study (Uther, 2004).

Loopholes are common and consequential in human society and culture. Yet, we know little about how this behavior emerges and develops. There are many children’s stories that play with the ambiguity between the letter and spirit of language, showing the humorous side of genuine misunderstandings and suggesting that children can appreciate the nuance of intentional ones. As just one of many examples, consider Peggy Parish’s ‘Amelia Bedelia’ series about a house-keeper who ‘dusts’ the living room by scattering dust on it, and ‘dresses’ the chicken in lederhosen. Children’s own willful misunderstandings are documented anthropologically in their games of guile (Opie & Opie, 2001). Closer to

home, parental anecdotes about loopholes abound, such as when the senior author of this paper told their child, “It’s time to put the tablet down,” only to have their child put the tablet physically down on the table and continue watching their movie (Bridgers, Schulz, & Ullman, 2021). Empirically, there is work on the *opposite* of loophole behavior in childhood: behavior that follows the spirit, but violates the letter of a rule. This research shows that children, from age four to ten, become increasingly more forgiving of behavior that violates the letter of a rule (but not the spirit) compared to behavior that breaks both (Bregant, Wellbery, & Shaw, 2019). These observations and research suggest that loophole behavior is not solely within the purview of adulthood, raising the question of when and how humans learn to find these creative workarounds.

Here, we present the first systematic and detailed developmental study of human loophole behavior. Inspired by prior empirical and computational work with adults, as well as relevant work in cognitive development, we propose that loophole behavior relies on a linked set of social-cognitive capacities: pragmatic language understanding, Theory-of-Mind (including an understanding of one’s own goals, and the goals of other people), and trading off utilities in joint planning (Bridgers et al., 2023; Qian et al., 2024). As we discuss in greater detail below, while the drive and ability to help and understand other people emerges early (Gergely & Csibra, 2003a; Warneken & Tomasello, 2006), a deeper comprehension of goals and ambiguity in language is needed to exploit the under-specification of social interactions, meaning that loophole behavior may emerge later in childhood. We predict that, once children are able to represent loopholes, they will both recognize loophole behavior as a separate behavior from compliance and non-compliance, and as useful for navigating conflicting goals among social partners.

Though developmental research that directly studies loophole behavior is scarce, there is a good deal of research on the social and cognitive capacities that we believe underlie it. In the rest of the introduction, we consider the development of these capacities in more detail. Specifically, we examine in brief the development of the ability to cooperate, to represent other people’s goals, to understand social ambiguity, and to trade-off between

one's own goals and those of other people. These different constitutive parts lead us to the suggestion that the understanding and use of loopholes will begin to emerge in children around ages five to six and continue to develop across early childhood. Fully detailing each of the constitutive parts would fill up several books, so we restrict ourselves to examining them in relation to our particular focus on loopholes. We turn first to the overall desire to cooperate, which breaks down when an individual intentionally misunderstands another person's desire or goal.

The ability to cooperate effectively is foundational to human's success as a species (Boyd & Richerson, 2005, 2009; Bratman, 1992; Henrich, 2015; Tomasello, 2009). This ability emerges early: within the first two years of life, infants and toddlers are motivated and able to take actions in the world that help other people achieve their goals (Buttelmann, Carpenter, & Tomasello, 2009; Cortes Barragan & Dweck, 2014; Liszkowski, Carpenter, & Tomasello, 2008; Svetlova, Nichols, & Brownell, 2010; Warneken & Tomasello, 2006). The robustness and sophistication of helpful, cooperative behavior then increases across early childhood (Bridgers, Jara-Ettinger, & Gweon, 2020; Brownell, Svetlova, & Nichols, 2009; Dunfield, Kuhlmeier, O'Connell, & Kelley, 2011; Martin & Olson, 2013; Meyer, van der Wel, & Hunnius, 2016; Svetlova et al., 2010; Warneken, Steinwender, Hamann, & Tomasello, 2014).

To figure out how to help and cooperate, children need to reason about others' beliefs, goals, costs, and rewards. While more complex Theory-of-Mind reasoning continues to develop past age four (e.g., higher-order beliefs about other people's beliefs, Tomasello, 2018), the basic aspects of representing and reasoning about the goals of others are present in infancy (Gergely & Csibra, 2003b; Warneken & Tomasello, 2007). Infants expect agents to minimize costs and maximize rewards when pursuing goals (Gergely, Nádasdy, Csibra, & Bíró, 1995; Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016; Liu & Spelke, 2017; Liu, Ullman, Tenenbaum, & Spelke, 2017), and it has been proposed that they also consider helping as one agent maximizing another agent's utility (Hamlin, Ullman, Tenenbaum, Goodman, & Baker, 2013; Powell, 2022; Woo, Liu, Gweon, & Spelke, 2024; Woo & Spelke,

2023). Later in development, between four and seven years of age, children can use estimates of other people's expected utilities in more sophisticated ways, such as resolving pragmatic ambiguity (Jara-Ettinger, Floyd, Huey, Tenenbaum, & Schulz, 2020), and deciding what is most helpful to teach (Bridgers et al., 2020; Gweon, Shafto, & Schulz, 2018).

But, while helping and cooperating are critical in many situations, our own goals are not always aligned with those of other people, and the most prosocial or cooperative action is not always the one most personally desirable. When goals are not aligned, the decision of whether to cooperate involves weighing one's own utility against that of another. Indeed, the same toddlers who are motivated to assist a person in need are less likely to do so when helping is physically or materially costly (Sommerville et al., 2018; Svetlova et al., 2010). Four- and five-year-old children can reason about their own and others' relative abilities to effectively allocate tasks that vary in difficulty depending on whether they are cooperating or competing (Magid, DePascale, & Schulz, 2018). So, while infants and toddlers consider others' goals and their own goals when deciding whether and how to help, more complex reasoning and trading-off of one's own utilities with those of others emerges later in childhood. These findings suggest that the ability to not just directly refuse to cooperate but instead use clever work-arounds, such as intentional misunderstandings, will likely also emerge later.

Representing and trading-off other people's goals would be difficult enough even if those goals were transparent. But an added complication for social reasoning is that goals are opaque. People cannot directly transfer the concepts, goals, and mental states in their heads to other people, and instead use ambiguous utterances that require an observer to recover the intended meaning through inference about possible alternative meanings (Bates, 1976; Goodman & Frank, 2016a). A rich developmental literature is devoted to this understanding of language in context, which begins in infancy and undergoes substantial development throughout childhood (Bohn & Frank, 2019). Of particular relevance for the current work, many studies show that 4- and 5-year-old children often struggle with identifying relevant alternatives for an utterance, such as knowing that "all" is an alternative to

“some” for the utterance “Some of the children are sleeping”, leading them to draw incorrect implicatures from what is said (e.g., Barner, Brooks, & Bale, 2011; Huang & Snedeker, 2009; Noveck, 2001; Skordos & Papafragou, 2016). Irony, metaphor, puns, and sarcasm further complicate the process of honing in on an intended meaning (McGhee, Goldstein, et al., 1983; Winner, Levy, Kaplan, & Rosenblatt, 1988). However, it should be noted that challenges in pragmatic reasoning in early childhood do not imply a discontinuity in the development of reasoning abilities, and there are strong arguments for the suggestion that the basis of pragmatic reasoning is largely available to children even before their first birthday (Bohn & Frank, 2019; Stiller, Goodman, & Frank, 2015). Also, while children as young as five (and possibly earlier) can reject the literal meaning of an utterance in favor of the pragmatic one, the ability to understand communicative intent continues to develop into adolescence (Demorest, Silberstein, Gardner, & Winner, 1983). Given that loophole behavior entails representing an unintended, alternative interpretation of an utterance in addition to the intended one, we might expect this behavior to emerge around five to seven years of age.

In addition to the previous work on the development of different cognitive building blocks of loophole behavior, there has been computational work directly examining adult loophole reasoning and decision-making (Qian et al., 2024). The proposed framework is that of a utility-theoretic recursive social reasoning model. The model simulates the decision-making process of a person that first infers the likely goal of a person making a request from their utterances, then weighs the trade-off between the decision-making person’s own goals, the goals of the person making a request, and the potential cost of complying, not complying, or using a loophole (Qian et al., 2024). The model also involves recursive reasoning about what the person making the request may infer about the decision-making agent’s intent from their actions (i.e., are they cooperative or uncooperative, genuinely misunderstanding or intentionally misunderstanding?). All told, the model brings together several different components studied separately in development: pragmatic reasoning about intent from ambiguous statements (Barner et al., 2011; Bohn & Frank, 2019), the representation

of another person's goals (Jara-Ettinger et al., 2016), beliefs about other people's beliefs (Tomasello, 2018), and trading off utilities in joint planning (Magid et al., 2018).

The recent formal framework used to study loopholes in adults combines separate components that may develop along different timelines, but must all be in place for the successful use of intentional misunderstandings and workarounds. The empirical and theoretical research in cognitive development on these components leads to the suggestion that an understanding and use of loopholes might emerge around age five and continue to develop through middle childhood. Across four studies, we examine the development of loophole behavior in children, by first establishing its overall ubiquity, and then triangulating its representation and use by studying its evaluation, prediction, and generation. More specifically, in Study 1 (experience), we survey parents to gather reports about the emergence and prevalence of loophole behavior in naturalistic settings. In Study 2 (evaluation), we test whether children consider loophole behavior as a means to mitigate the costs of refusing the request of a person who has an opposing goal. In Study 3 (prediction), we test whether children anticipate that loopholes will be used more often when people have conflicting vs. aligned goals. Finally, in Study 4 (generation), we examine whether and at what age children can actively come up with loopholes themselves in response to a given directive. We conclude by discussing how these studies provide converging evidence for the developmental timeline of loophole behavior in childhood, the role of humor and deniability in evaluating loopholes, and potential limitations given possible cultural differences.

### **Study 1: Children's real-life experience with loopholes**

Given the scarcity of data available on the use of 'loopholes' by children, we began our investigation by surveying U.S.-based parents about their children's tendency to exploit loopholes in their daily lives. This survey serves as a starting point to begin to characterize the emergence, extent, and scope of this behavior in childhood.

## Methods

### Participants

Participants were 260 parents of children between the ages 3 and 18 years (inclusive), recruited online via Prolific in October 2020. The survey took approximately 9 minutes to complete, and the compensation was \$1.43. Participants were U.S. residents, fluent in English, and from diverse geographical regions and educational backgrounds. Participants reported on 425 children in total ( $M_{age}$ : 8.7, range: 3 to 18 yrs; 42% female, 5% declined to state) from diverse ethnic and cultural backgrounds (34% White, 10% multiracial, 4% Black, 3% Asian, 3% Hispanic or Latinx, 47% declined to state). An additional 39 participants were recruited but excluded from analysis due to failing the comprehension check (N = 7), or not having children of a relevant age (N = 32).

### Procedure

The survey was implemented using Qualtrics (Fig. 1) . In the introductory phase of the survey, participants (parents) were first given a definition of loophole behavior (“Children (and adults) may understand the actual intended meaning of what was said to them or asked of them, but choose to interpret things differently.”), as well as examples of children finding loopholes in parents’ directives, and examples of non-loophole behavior (i.e., direct ignoring or refusing, and genuine misunderstanding). Parents then indicated whether they understood what was meant by loophole behavior (first comprehension check). To further clarify the meaning of loophole behavior and test understanding, parents were presented with two stories and asked to classify loophole vs. non-compliant behaviors with feedback (comprehension quiz). After these stories, parents were again asked to indicate whether or not they understood the concept of loopholes (second comprehension check).

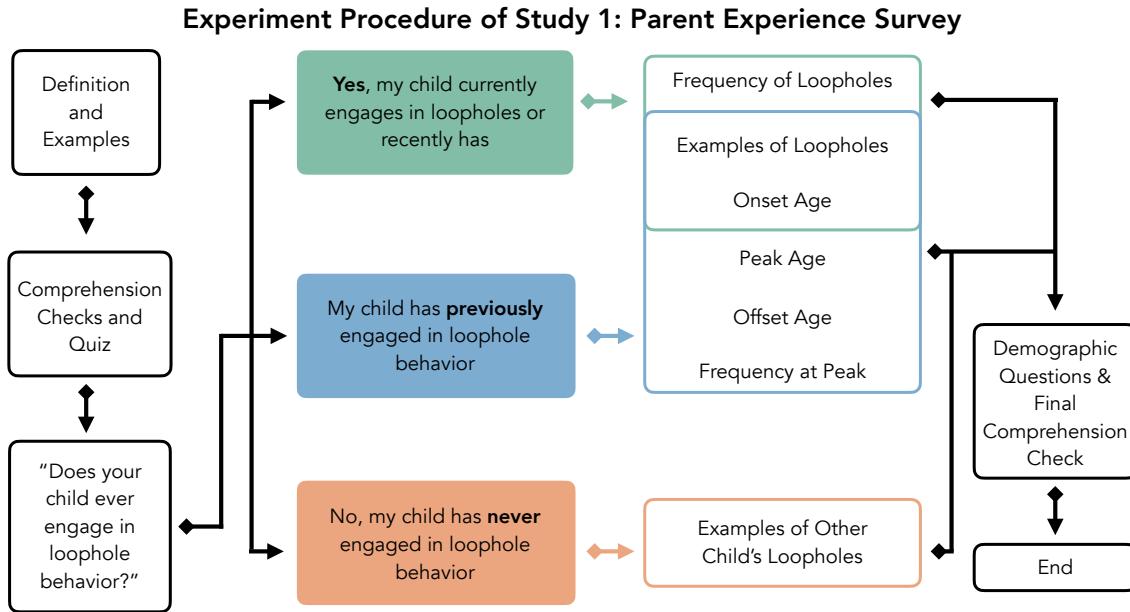
In the main phase of the survey, parents were asked to report for each of their own children: (1) the child’s current age, and (2) whether they *currently* engage, *used to* engage, or *have never* engaged with loopholes. Parents who said that their children currently engage or previously engaged with loopholes were then asked to provide the age of onset

(i.e., when their children first began engaging with loopholes), and in addition, for children who previously engaged, parents were asked to provide the age at which loophole behavior was at its peak, and the age of offset (i.e., when their children stopped using loopholes). Parents of both children who currently or previously used loopholes were also asked how frequently this behavior occurred on a 5 point scale (i.e., “several times a day”, “about once a day”, “once every few days”, “once every few weeks”, “less frequently than once every few weeks”). Finally, these parents were invited to share examples of their children’s loophole behavior. Parents who said that their children never engaged with loopholes were invited to share anecdotes of other children’s loophole behavior (if applicable). For all parents, the survey ended with a series of demographic questions, as well as a final comprehension check where they had to describe what they thought the study was about.

## Results and Discussion

Overall, parents indicated that they readily understood what was meant by loophole behavior. In the comprehension quiz where parents were asked to classify a child’s behavior in response to a parent’s directive in two stories, 93% correctly identified loophole behavior and 91% correctly identified non-compliance. In addition, after this quiz, on the second comprehension check, 98% of parents indicated that they understood what was meant by loophole behavior, while only 1.9% indicated that they “Maybe” understood and only one parent indicated that “No”, they did not understand.

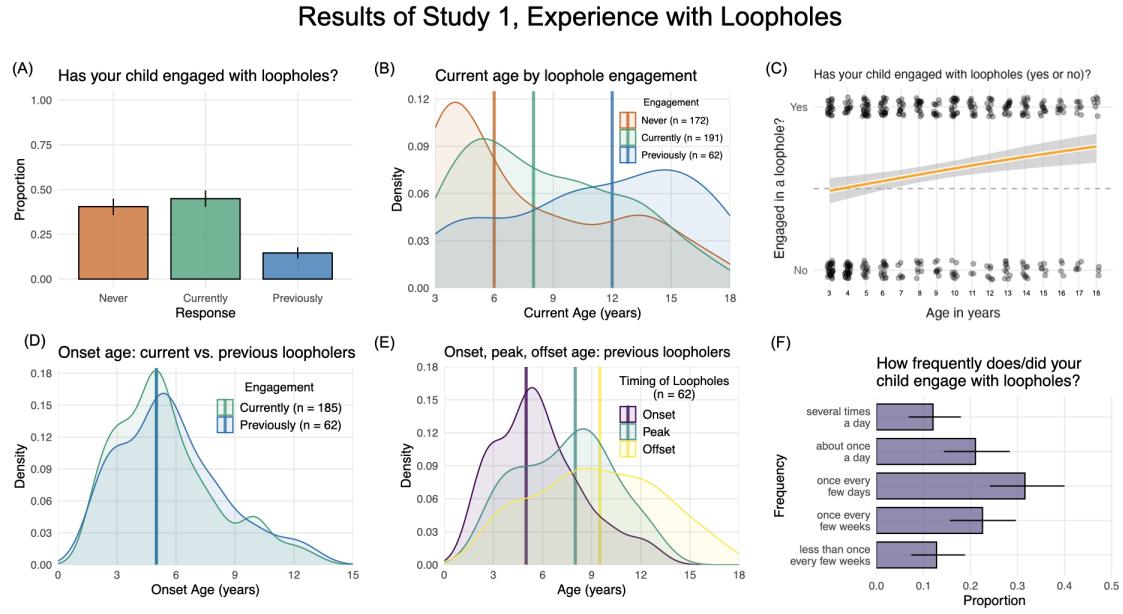
We found that loophole behavior is common in parent-child interactions. A majority of children (60%; N = 253) were reported as engaging in loophole behavior currently (44.9%, 95% CIs: [40.2%, 49.9%]; N = 191) or previously (15%, 95% CIs [11.4%, 17.9%]; N = 62) (Fig. 2(A)). Looking at the distribution of children’s ages in each of these groups (Fig. 2(B)), shows that children who were reported as never having engaged in loopholes are younger than their counterparts (modal age is around four years, and the median is six years), children who are currently engaging in loopholes are on average older (modal age around five years and median eight years), and children who have previously engaged in loopholes



*Figure 1.* Survey Structure, Study 1. Following an introduction, participants indicated if their child ever engaged in loophole behavior, and answered a set of questions based on whether their child currently engages, previously engaged, or never engaged in loophole behavior. Parents of children who currently engage in loopholes were asked about Onset Age (when this behavior began) and Frequency (how often their child engaged in this behavior). Parents of children who previously engaged were also asked for Onset Age, frequency of the behavior when it was at its peak, and Offset Age (when the behavior stopped). All parents were asked for examples of their children’s loophole behaviors (if parent’s children had never engaged with loopholes, they were invited to share examples of other children’s loopholes).

but no longer do are older still (modal age around 15 and median age 12). We also observe a switch between the majority of children reported as ‘never’ having engaged with loopholes switching to ‘currently’ occurring between four and six years of age, and a switch from a majority of ‘currently’ to ‘previously’ from six to 12.

To test if there was indeed a significant correlation between current age, and whether or not children were reported as ever having engaged with loopholes (i.e., either currently or previously), we fit a Bayesian logistic regression predicting children’s binary loophole engagement ('yes' for 'currently' and 'previously' and 'no' for 'never') from a fixed effect of children’s current age in months (continuous and centered) (*binary\_engagement ~ age\_centered*). As can be seen in Fig. 2 (C), this analysis revealed a significant effect of



*Figure 2.* Study 1 Results. (A) Proportion of children ( $N = 425$ ) who have never engaged (orange), are currently engaging (green), or have previously engaged (blue) with loopholes; a majority of children (60%) are in the ‘currently’ or ‘previously’ categories. Error bars are 95% bootstrapped CIs of the mean. All bootstrapped CIs displayed in the figures of this paper were calculated using the R-package `tidyboot::tidy_mean`. (B) Distribution and median current age for children who have never engaged (orange; median = 6 years), are currently engaging (green; median = 8 years), or have previously engaged (blue; median = 12 years) with loopholes. (C) Whether or not children have ever engaged with loopholes (i.e., ‘yes’: ‘currently’ and ‘previously’; ‘no’: ‘never’) by current age in years with fit model line in orange from R-package `ggplot::geom_smooth` using method “`glm`”; dots are individual parent responses per child. (D) Distribution and median age of onset for loophole behavior for ‘currently’ (green; median = 5 years) and ‘previously’ (blue; median = 5 years). (E) Distribution and median age of loophole onset (purple; median = 5 years), peak (turquoise; median = 8 years), and offset (yellow; median = 9.5 years) for ‘previously’ ( $N = 62$ ). (F) Proportion of current and previous engagement with loopholes by level of frequency; mode is “once every few days”. Error bars are 95% bootstrapped CIs of the mean.

age: As age increased, children were more likely to have engaged in loophole behavior, either currently or previously ( $\beta = 0.007$ , 95% CI: [0.003, 0.011]). (The R-package `brms::brm` was used to run all analyses reported in this paper.) As a decision rule for comparisons here and throughout, we see if the 95% confidence interval for the parameters of the relevant contrast includes 0. Note that for Bayesian regression models, the parameters can be interpreted in the same way as the parameters of frequentist regression models. The only difference is

that for the Bayesian regression model the 95% confidence intervals can be interpreted as there being a 95% chance that the parameter falls within that interval, as opposed to the different, less intuitive frequentist interpretation of confidence intervals (i.e., that 95% of the possible confidence intervals contain the true parameter; see Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2016, for details).

We also found that the ages parents reported for onset, peak, and offset of loophole behavior line up with the distribution of current age split by loophole engagement (i.e., ‘never’, ‘currently’, ‘previously’). As shown in (Fig. 2(E and D)), parents of children who currently or previously engaged with loopholes reported that their children began engaging with loopholes at five-to-six years of age on average ( $M_{age}$ : 5.6 years, range: 2 to 13 years), and the frequency of parents reporting that their child has ‘never’ used loopholes declines from ages three to six, as ‘currently’ responses increase in frequency (Fig. 2(B)). Regarding peak and offset age, parents of children who previously engaged in loopholes ( $N = 62$ ) reported that their children engaged with loopholes most frequently at ages seven to eight ( $M_{age}$ : 7.4 years, range: 2 to 13 years), and that loophole use tapered off around ages nine to ten ( $M_{age}$ : 9.3, range: 3 to 17 years) (Fig. 2(E)). Again, these results line up nicely with the distributions of parents’ responses to the loophole engagement question: ‘currently’ responses gradually decrease from 6 to 18, while ‘previously’ responses gradually increase until around age 15, with cross-over around age ten or eleven (Fig. 2(B)). Given that these are different subsets of parents answering different questions, we take this triangulation to suggest that the parental reports collected have internal reliability.

The results discussed so far indicate that a majority of children have engaged with loopholes at least once, and offer a potential developmental trajectory for when this behavior emerges, reaches its peak, and declines in frequency. We additionally asked parents of current loopholers how frequently their children engaged in this behavior and of previous loopholers, how frequent the behavior was at its peak. Responses to this question were roughly normally distributed across the scale provided, with the modal response indicating that children who engage in this behavior do so regularly, the mode being “once every few

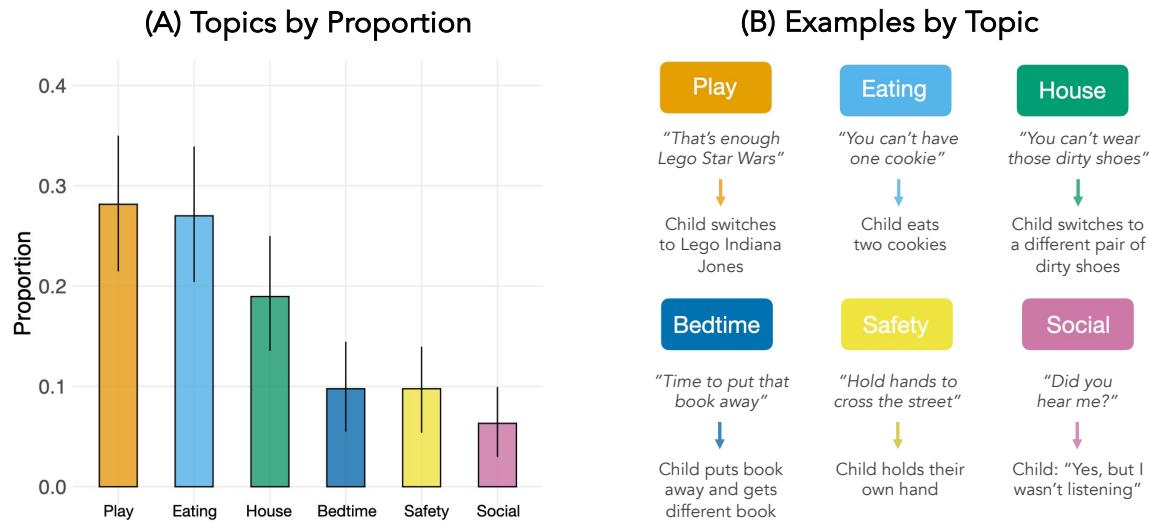
days” (Fig. 2(F)).

Beyond reporting about age and frequency, parents readily shared examples of their own (or other people’s) children’s loophole behavior. Parents who said that their children currently or previously used loopholes collectively shared 268 examples. The majority of these (64.9% (95% CIs: [59.6%, 70.2%])) were validated as ‘loophole behavior’, followed by ‘other’ (14.2%; 95% CIs: [10.3%, 18.3%]) and ‘non-compliance’ (13.1%; 95% CIs: [9.2%, 17.4%]). (See the Supplementary Materials for more details.)

The examples parents shared spanned linguistic utterances and behavioral domains. Regarding different linguistic utterances, children found loopholes in directives having to do with *reference* (a child holding candy was told they couldn’t eat “that candy” so they ate a different piece of candy), *timing* (a child was told they needed to be home by seven, and they came home at 7am the next day), *number* (a child was told they could not have one cookie, so they had two), *scope of generalization* (a child was told to stop playing their Lego Star Wars video game, so they switched to their Lego Indiana Jones game), *scalar terms* (a child was told to get “some sunlight”, so they went outside for a second and then came back inside), *indirect requests* (a child was asked if they could stop playing, and they said, “Yes, I can,” only to continue to play), and more.

To explore the different domains children’s loopholes fell into, the first authors reviewed the examples and identified a set of six domains that captured their diversity: 1) *rules around play*: when and how it should be done; 2) *rules around eating*: what, when, and how to eat; 3) *rules around the house*: how to behave indoors, as well as chores and responsibilities; 4) *rules around safety*: how to be safe and personally hygienic; 5) *rules around bedtime*: bedtime routines and timing; 6) *rules around social interaction*: how to behave well with others, as well as how to treat pets and animals (See Fig. 3). The first authors treated these categories as mutually exclusive and classified each loophole example into one of these domains ( $\kappa = 76.8\%$ ; if a behavior could fall under two or more domains, it was classified by what was considered most central). (See Supplementary Materials for more details and examples of this coding.) Parents reported loopholes predominantly having to

### Study 1, Parent Report: Topics of Children's Loopholes



*Figure 3.* Study 1, Parent report of loophole examples *Parents shared examples of children exploiting loopholes across a variety of domains, testing the boundaries of rules around play (orange), eating (light blue), the house (green), bedtime (dark blue), safety (yellow), and social interactions (pink) with people and animals.* (A) Proportion of loophole examples that involved a given domain. Error bars are 95% bootstrapped CIs of the mean per domain. (B) Representative examples by topic.

do with rules around play (28.2%, 95% CIs: [21.8%, 34.9%]) and around eating (27.0%, 95% CIs: [20.8%, 33.5%]). Within play, the majority of loopholes involved screen and technology usage (e.g., a child is told to stop playing on the computer, so they switch to playing on an iPad), and within eating, the majority of loopholes involved eating sweets (e.g., a child is told no more gummy bears, so they switch to eating gummy worms).

In summary, according to parent report, loophole behavior is prevalent and frequent in children's everyday interactions with their parents. The linguistic and behavioral diversity of the loophole behaviors shared suggests that children's ability to find loopholes in linguistic utterances is a general cognitive phenomenon, rather than being specific to the emerging understanding of a particular linguistic construction, or emerging experience with particular conceptual or behavioral domains. These anecdotes begin to paint a picture of the constellation of rules and norms about which young children are learning and then testing the boundaries of. These findings expand prior work showing that adults regularly

engage in loophole behavior in their everyday social interactions to provide evidence that this behavior begins in childhood (Bridgers et al., 2023).

The results of the survey also provide a hypothesis for when loophole behavior emerges and the trajectory of its understanding and use across early to middle childhood. Converging evidence across different questions examining the ages of children who currently engage with loopholes and children who previously engaged, including an explicit question about the age of onset, suggests that loophole behavior emerges around five to six years of age. This age of onset coincides with what we predicted based on prior developmental literature on potentially related social-cognitive abilities such as pragmatic implicature (Barner et al., 2011; Bohn & Frank, 2019), naive-utility calculus (Jara-Ettinger et al., 2016), and higher-order Theory-of-mind (Tomasello, 2018). Interestingly, one might have expected parents of children who previously engaged in loopholes to have a fuzzier memory of when their children first began using loopholes compared to parents of children who are currently engaging in them. The distributions of reported age of onset for these two groups of parents, however, are nearly identical (Fig. 2(E)), suggesting that initial observance of this behavior may be quite memorable for parents. Parents of previous loopholers additionally reported that this behavior reaches its peak around seven to eight years of age, and then tapers off around nine to ten years and into adolescence.

Study 1 established loopholes as an ecologically valid behavior in childhood, and shows parents can easily distinguish it from compliance, defiance, and confusion. However, it relied on parent report, and there is a need to establish the validity of these findings with children themselves. While the parents' responses were internally consistent, they depend on children actually exploiting loopholes with their parents, and non-expert intuitions about what is observed. It's possible children may understand loopholes but not use them, and even if they do use them, parents may not catch the first instances. It may be that children are able to understand loopholes far earlier than parents start noticing their use. Further, Study 1 leaves open the question of the purpose of children's engagement with loopholes. Previous work with adults found that adults consider loopholes to be less costly

than outright non-compliance, suggesting that loopholes may be used as a tool to mitigate consequences when one does not want to comply (Bridgers et al., 2023; Qian et al., 2024), but it remains to be seen whether children understand loopholes in this way too. The following experiments aimed to examine children's understanding and use of loopholes by studying children directly.

### **Study 2: Children's evaluation of loopholes**

We empirically tested whether children view loopholes as less costly than non-compliance by examining whether 4- to 9-year-olds estimate that loopholes will result in less punishment than non-compliance. In a pre-registered experiment, we presented participants with stories of children complying, not complying, or finding a loophole in a parent's directive, and asked participants how much trouble the child protagonists would get into. In addition, given that adults find loopholes humorous (Bridgers et al., 2023; Qian et al., 2024), we explored whether children also find loopholes more amusing than non-compliance or compliance by recording whether they laughed or smiled upon learning how the child protagonist responded. We selected four years (48 months) up until ten years (121 months) as our age range so that it began a bit before the age of onset (five to six years) and went up until the age of offset (nine to ten years) for loophole behavior, as reported by parents in Study 1.

We predicted that, like adults, children would think that loophole behavior incurs less trouble (or results in less punishment) than non-compliance. We also predicted that children would find the loophole response funnier than either non-compliance or compliance. Also, given parent reports of loophole engagement and prior literature about possibly related social-cognitive abilities, we anticipated that we would likely observe developmental differences in children's evaluations of loopholes compared to compliance and non-compliance. However, as discussed it is possible that loophole comprehension may precede loophole production, and though parents don't often observe loopholes before age five, it is possible that children may understand and differentiate loopholes from other behaviors before then.

If this were the case, then we might not observe differences in children's evaluations of loopholes across the age range studied.

If we were to find developmental differences, though, there are a-priori at least two plausible patterns we might observe. These patterns depend on whether young children see loophole behavior as more similar to compliance, or to non-compliance. On the one hand, perhaps 4- and 5-year-olds have trouble understanding the actual intent of a parent's directive, and evaluate loophole behavior as truly fulfilling the request. If so, then younger children would be more likely to rate loophole behavior as similar to compliance. On the other hand, perhaps younger children are able to recover a parent's intended meaning, and recognize that the loophole violates this intent (while having more trouble than older children at keeping in mind alternative, less plausible interpretations of the utterance). In such a case, younger children would be more likely to rate loophole behavior as similar to non-compliance. Preliminary related work suggests that the latter option was more likely (Bridgers et al., 2021), and so we predicted that children would increasingly rate loopholes as resulting in less trouble from age four to ten years, and that the difference between loopholes and non-compliance would also increase.

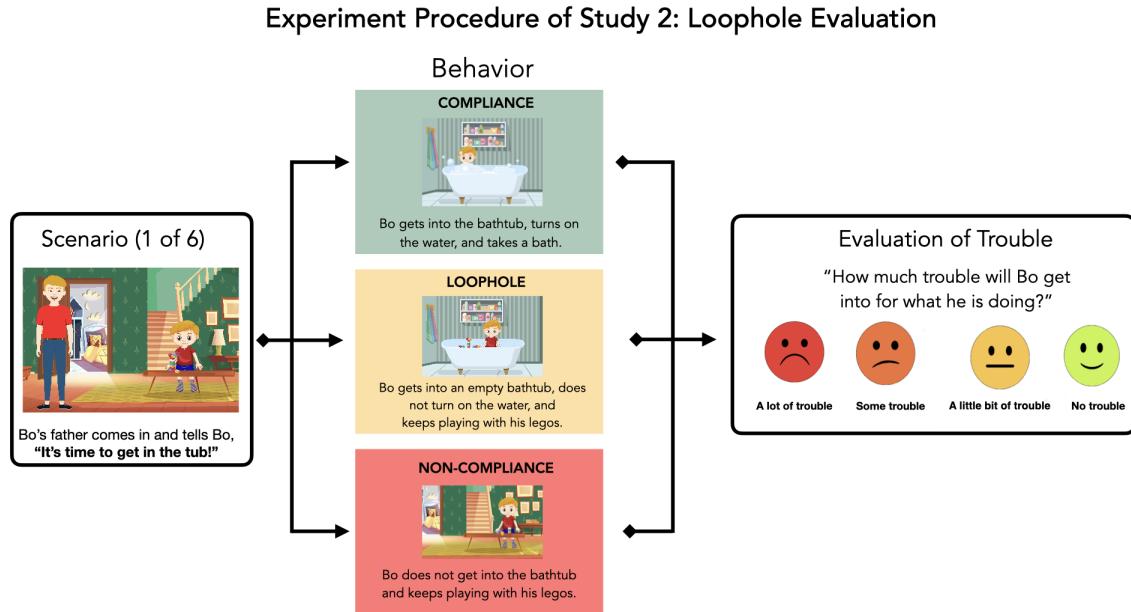
## Methods

**Participants.** We recruited 108 4- to 9-year-olds ( $M_{age}$ : 7.07 yrs, range: 4.07 to 10.10 yrs; 51% female; 74% White, 11% Asian, 5% Hispanic, 4 % Other, 5% Mixed, 1% American Indian, 1% did not report) via Children Helping Science (an online platform where researchers can post studies and families can sign up to participate). Children participated asynchronously in a self-moderated experiment hosted on the platform from October 2022 to August 2023. To be included in analysis, children had to be between 48 and 121 months of age (inclusive), not have participated in a prior related study, be fluent in English, and pass a set of inclusion criteria, which were a series of comprehension checks that assessed whether children understood the trouble scale and recognized that the compliant behaviors followed the parents' directives and so were not deserving of trouble (see Procedure for

details). An additional 26 participants were recruited but excluded from analysis, due to failure to meet inclusion criteria ( $N = 11$ ), having previously participated in the study or a closely related study ( $N = 4$ ), not having video turned on for the study so we were unable to verify if a child was present ( $N = 8$ ), or parental interference during testing ( $N = 3$ ). The sample size, inclusion and exclusion criteria, as well as the hypotheses and analyses were all preregistered on the Open Science Framework.

**Procedure.** Participants were told that they were going to hear stories about children and their parents, and that in each story the experimenter would need their help to figure out how much trouble the child would get into for what they were doing. The stimuli were presented as novel story-books, and narrated by an experimenter. We developed twelve scenarios based on the real-world examples of loopholes parents shared in Study 1. In each scenario, a parent gave a directive to a child (phrased as a request or demand), and the child then responded in one of three ways: (1) compliance, (2) non-compliance, or (3) loophole. The loophole was the exact or a slightly modified version of the real-world loophole, and we designed the corresponding compliant and non-compliant behaviors. Each participant saw six of the twelve possible scenarios. The conditions of the scenarios were counter-balanced so that in two scenarios the child protagonist engaged in a loophole, in two scenarios the protagonist complied, and in two scenarios the protagonist refused to comply. Which six of the twelve scenarios participants saw and the condition of each scenario (i.e., compliance, non-compliance, or loophole) were randomized across participants. The order of the conditions was pseudo-randomized, such that participants viewed one of each condition (compliance, non-compliance, loophole) in the first three trials and one of each condition in the last three trials. The order within each set of three trials was randomized across participants.

For each story, after learning how the child protagonist responded to the parent's directive, child participants were asked how much trouble the child protagonist would get into (see Fig. 4). Children indicated the amount of trouble by selecting a rating along a 4-point scale from "a lot of trouble" to "no trouble", where each point was represented by a



*Figure 4.* Structure of Study 2 (Evaluation). Children aged four to nine were presented with stories about parents and their children, in which a parent gave a directive to their child. In each story, the child protagonist could either comply with, not comply with, or find a loophole in the directive. Child participants were asked to rate how much trouble the child protagonist would get into, using a four-point Likert scale represented by verbal and written labels and different colored emoji faces ranging from a red frowning face ("a lot of trouble") to a green smiling face ("no trouble"). Each child was presented with 6 (of 12) scenarios that counterbalanced compliance, non-compliance, and loophole behavior (i.e., children saw two of each behavior).

different colored emoji-face ranging from frowning (red) to smiling (green). Children could select the amount of trouble by pointing to a face and having their parent click on it, or clicking on the face themselves. Children received training and practiced using the trouble scale ahead of time. They also received three training trials in which a child should get into a lot of trouble (ripping their sister's blanket and pushing her), a little bit of trouble (accidentally breaking a flower vase), and no trouble (helping their brother build a block tower). To be included in analysis, children needed to rate (i) the ripping and pushing story as resulting in either getting into "some trouble" or "a lot of trouble" (the two higher levels of trouble on the scale), (ii) the helping story as resulting in "no trouble" (the lowest level of trouble on the scale), and (iii) the two test trials where a child complies with a parent's

request as resulting in either “a little bit of trouble” or “no trouble” (the two lower levels of trouble on the scale). In the final test trial, children were asked to explain their choice of trouble. As an exploratory measure, we also coded children’s own amusement and affect upon hearing the child protagonist’s response in the six test trials (indexed by whether they smiled or laughed).

## Results and Discussion

We predicted that there would be a main effect of condition such that children would rate loophole behavior as getting the protagonist into *more* trouble than compliance but *less* trouble than non-compliance. We also predicted that if there were an effect of age, it would likely be an interaction with condition where children’s ratings of trouble for loophole behavior (and perhaps non-compliant behavior) would change with age, but ratings of compliant behavior would be low, and not vary with age.

To test our hypotheses, we fit a pre-registered, confirmatory Bayesian mixed effects cumulative logit model (a Bayesian ordinal regression) predicting children’s ratings of trouble (4-level factor from 1 to 4) from fixed effects of condition (3-level factor: loophole, non-compliance, compliance, with loophole as the reference category) and age in months (continuous and centered) with maximal random effects (random intercepts and effects of condition by subject and by story/scenario). In addition to this additive model, we fit a simpler model with a single predictor of condition (no age), and a more complicated model with fixed effects of condition, age, and their interaction. Formal model comparison preferred the additive model over the other two models, so we report the results of the additive model here (i.e.,  $rating_{trouble} \sim condition + age_{centered} + (1 + condition|subject) + (1 + condition|story)$ ). Model comparisons here and throughout were conducted using expected log-posterior densities, and we report the outcomes of model comparison in terms of (posterior) model weights (McElreath, 2018; Nicenboim, Schad, & Vasishth, 2021). Please see Supplementary Materials for more details on these comparisons.

As shown in Fig.5(A), this analysis found a main effect of condition, such that children

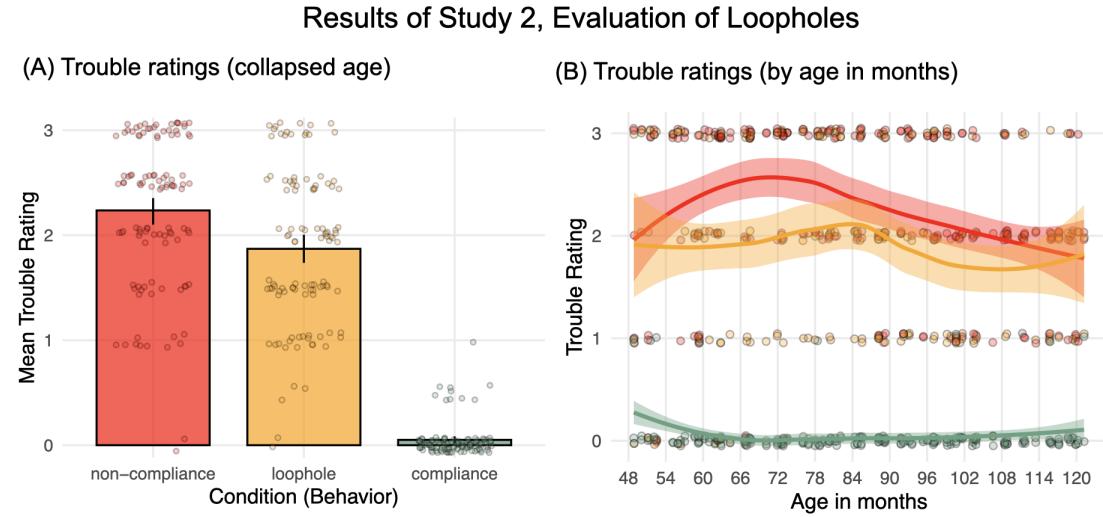


Figure 5. Results of Study 2: Loophole Evaluation (Trouble). (A) Children's ( $N = 108$ ) mean ratings of trouble on a 4-point scale from 0 to 3 for non-compliance (red left-bar), loopholes (yellow middle-bar), and compliance (green right-bar). “No trouble” is 0 and “A lot of trouble” is 3. Dots are individual subject mean ratings (two responses per subject per condition averaged for 108 total ratings per bar), and error bars are 95% bootstrapped CIs of the subject means. (B) Ratings of trouble by age in months per condition; dots are individual subjects (two per subject per condition for 648 ratings in total) with fit model lines by condition from R-package `ggplot::geom_smooth` using method “loess”.

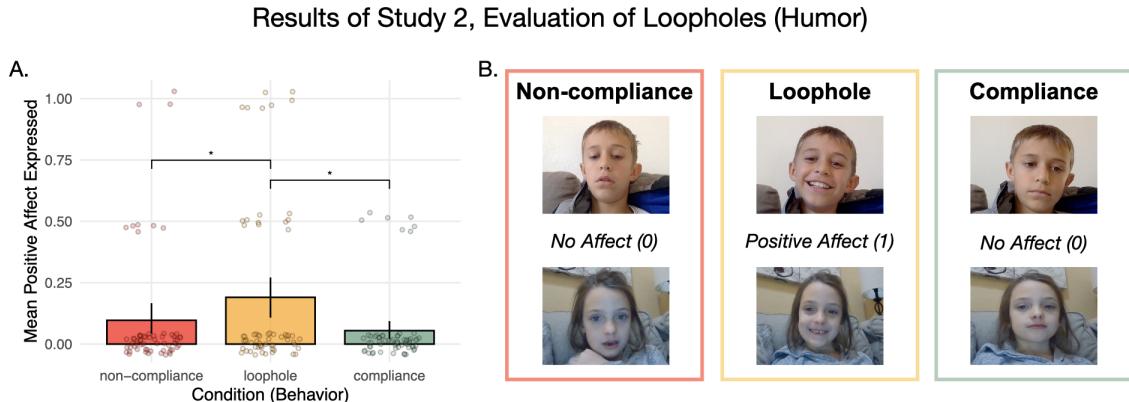
rated loopholes as getting the protagonist into more trouble than compliance ( $\beta = 9.07$ , 95% CI: [6.71, 12.41]) but less trouble than non-compliance ( $\beta = -0.99$ , 95% CI: [-1.55, -0.48]). There was also a main effect of age such that as age increased, children’s ratings of trouble for loopholes decreased ( $\beta = -0.02$ , 95% CI: [-0.035, -0.004]). Refactoring the condition so that non-compliance was the reference category and refitting the same model revealed that children’s ratings of non-compliance also decreased as age increased ( $\beta = -0.03$ , 95% CI: [-0.06, -0.01]) (See Fig.5(B)).

To explore whether children found loophole behavior more amusing than non-compliance or compliance, we coded their emotional reactions (indexed via their facial expression) upon hearing how the child protagonist responded to the parent’s directive in each story. For all children for whom their entire face was visible from the video recording ( $N = 65$ ), we coded whether the child laughed or smiled as ‘1’ and ‘0’ otherwise (see Supplementary Materials for more details on this coding). We next fit an exploratory Bayesian

mixed effects logistic regression predicting children's affect (0 or 1) from fixed effects of condition (3-level factor: loophole, non-compliance, compliance, with loophole as the reference category) and age in months (continuous and centered), with random intercepts and effects of condition by subject and story/scenario. In addition, we again fit a simpler model with the single predictor of condition, and a more complex interactive model. Formal model comparison preferred the additive model, so we report the results of this model here ( $\text{positiveaffect} \sim \text{condition} + \text{age}_{\text{centered}} + (1 + \text{condition}|\text{subject}) + (1 + \text{condition}|\text{story})$ ). As shown in Fig. 6(A), while children generally did not laugh or smile on a majority of trials, the analysis indicated that children expressed positive affect significantly more often when hearing of a child engaging in loophole behavior, compared to both non-compliance ( $\beta = 1.95$ , 95% CI: [0.39, 4.59]) and compliance ( $\beta = 3.02$ , 95% CI= [0.99, 6.90]). Children's tendency to find loopholes funny did not significantly increase with age ( $\beta = 0.03$ , 95% CI = [-0.01, 0.07]). Children's tendency to find non-compliance funny, however, did significantly decrease with age ( $\beta = -0.03$ , 95% CI = [-0.06, -0.01]). We are hesitant to interpret these age effects as the instances of positive affect were quite low (children only laughed or smiled on 39 out of 356 trials).

In sum, as predicted, we found that when collapsing across age, children, ages four to ten years, evaluated loopholes as leading to more trouble than compliance, and less trouble than non-compliance. We also found that children considered loopholes funnier than either compliance or non-compliance, providing further evidence that they distinguish loopholes from both of these behaviors.

We also predicted that if we found an effect of age it would likely be an interaction such that across the age range studied children's ratings of trouble for compliance and non-compliance would stay relatively constant, but their ratings of trouble for loopholes would decrease. Instead, however, formal model comparison preferred a model with additive rather than interactive effects of condition and age. This analysis indicated that children's ratings of trouble for both loopholes and non-compliance decreased with age, but that the relative difference in ratings for each behavior stayed more or less constant. That being



*Figure 6.* Results of Study 2: Loophole Evaluation (Humor). (A) Mean positive affect expressed for non-compliance (red left-bar), loopholes (yellow middle-bar), and compliance (green right-bar). Children ( $N = 65$  participants, 356 total trials) were coded as 1 if they laughed or smiled, and 0 otherwise upon hearing how the child protagonist responded to the parent’s directive. Error bars are 95% bootstrapped CIs of the subject means (means are calculated first by subject and then by group per condition). (B) Example images of children’s reactions to the child protagonist’s response by condition (behavior), and the related affect coding.

said, a closer inspection of children’s trouble ratings by age, as can be seen in Fig. 5(B), suggests that this difference between non-compliance and loopholes may expand and then contract with both 4-year-olds and 9-year-olds rating the two behaviors similarly, and 5- to 8-year-olds rating loopholes as resulting in less trouble than non-compliance. We should be cautious to not over-interpret this pattern of data, as additional exploratory analyses to see if there was evidence for a parabolic effect of age were inconclusive and formal model comparison preferred the additive over the interactive model (see Supplementary Materials for more details). But, we note that this pattern interestingly mirrors parents’ reports in Study 1 of how their children’s loophole production emerged around five to six years of age, peaked around seven to eight, and decreased around nine to ten, suggesting that children’s tendency to engage with loopholes may depend on how likely they think the behavior will reduce the probability or severity of trouble. Future research over-sampling children at the ends and middle of this age range, as well as beyond the age of 10 years could confirm whether a linear or parabolic effect of age best captures children’s evaluations of loopholes vs. non-compliance in terms of trouble.

Study 2 provides initial evidence that children consider loopholes as a distinct behavior from non-compliance, and that this appreciation may increase across early and middle childhood. The results also provide insight as to *why* children might exploit loopholes (i.e., to get out of trouble), in line with findings from adults (Bridgers et al., 2023; Qian et al., 2024). Indeed, children's lower ratings of trouble for loopholes compared to non-compliance are also consistent with adults' ratings of these behaviors: in a separate study, we asked adults to evaluate similar parent-child interactions and rate the trouble these behaviors would incur, how upset the parent would be, and how funny the parent would find them. Adults estimated that children's loophole behavior would incur less trouble and upset than non-compliance and more humor than either non-compliance or compliance. Children's responses in Study 2 thus seem to reflect the true state of affairs and may reflect their own experience of getting out of trouble via a loophole. (See Supplementary Material for details on this study with adults and full results.)

In Study 3, we continue the direct testing of children, and examine an additional part of our hypothesis: that children may use loopholes as a way to get around conflicting goals.

### **Study 3: Children's predictions of others' use of loopholes**

We investigated *when* children predict loopholes will be used, by manipulating goal alignment between parents and children in social interactions (i.e., the child protagonist's goals were either in agreement or at odds with their parent's goals). We then asked children to predict which behavior (compliance, non-compliance, or loopholes) would be used by the child protagonist in a given scenario in response to the parent's directive. We selected a slightly older age range (five up until ten years) since piloting suggested that younger children struggled with the highly verbal and three-alternative force-choice structure of the task.

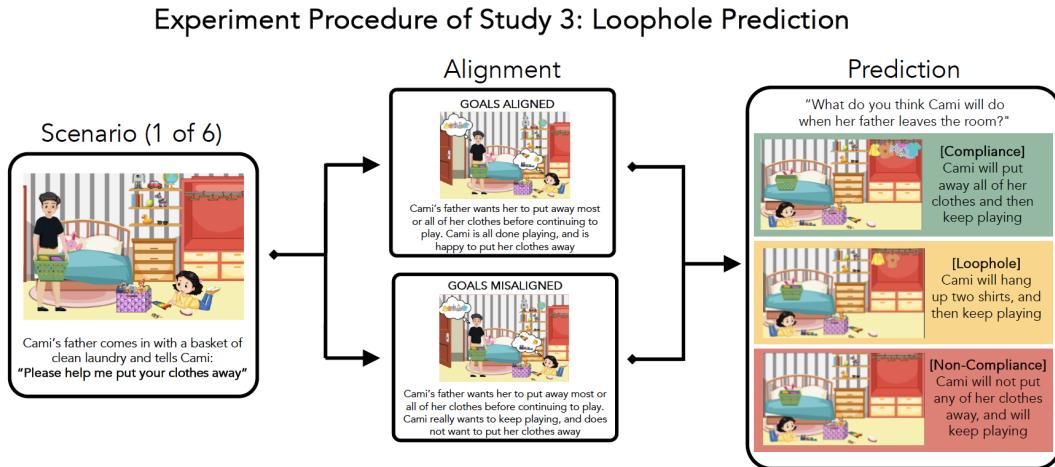
Given previous work with adults (Bridgers et al., 2023), we hypothesized that children would be more likely to predict compliant behavior when goals were aligned, and more likely to predict non-compliant behavior when goals were misaligned. More importantly for our

focus, we also predicted that children would expect loopholes to be used more often in cases where goals were misaligned. Given the emerging pattern of loophole behavior and evaluation in Studies 1 and 2, one possible developmental pattern that we might expect to see is that as children get older, their tendency to predict loopholes more often when goals are misaligned (than aligned) would increase.

## Methods

**Participants.** We recruited 140 5- to 9-year olds ( $M_{age}$ : 7.54 yrs, range: 5.03 to 10.38 yrs; 49% female; 54% White, 19% Asian, 14% Mixed, 5% Latinx, 3% Black, 1% Middle Eastern, 4% did not report) online again through the Children Helping Science platform. Children participated asynchronously in a self-moderated experiment from January to September 2023. An additional 42 participants were excluded from analysis due to having previously participated in the study or a closely related study ( $N = 35$ ) (we couldn't restrict study access based on prior study participation), parental interference during testing ( $N = 2$ ), or participating after we reached the total pre-registered number of children ( $N = 5$ ). The sample size, inclusion and exclusion criteria, as well as the hypotheses and analyses were all preregistered on the Open Science Framework.

**Procedure.** We used the same twelve scenarios used in Study 2, in which a parent issues a directive to a child. However, this task differed in that after the directive was given (e.g., “Help me put your clothes away.”), participants heard additional information about the goals of the parent and child. Specifically, they either heard that the parent’s and child’s goals were aligned, or that they were misaligned (e.g., the line “Cami’s father wants her to put away most or all of her clothes...” is followed by either Cami “...is happy to put her clothes away” or “...Cami really does not want to put her clothes away”). Following each scenario, participants were asked to predict what the child protagonist would do next, choosing among three possible behaviors: compliance, non-compliance, or loophole. These labels were not used when presenting the behavior options to children, rather the behavior itself was described. Pictures depicting each behavior were displayed at the same time



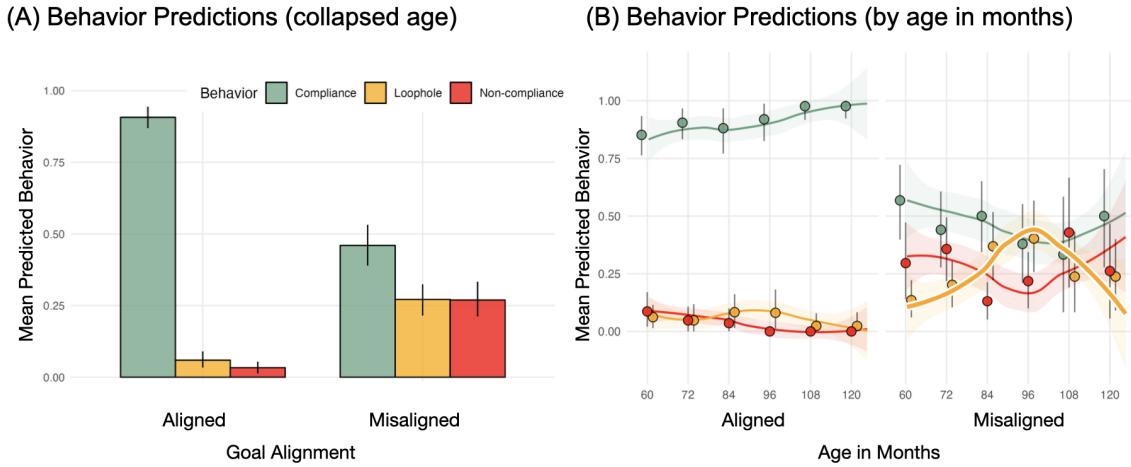
*Figure 7. Structure of Study 3 (Prediction). Children were presented with stories in which a parent gave their child a directive. Participants then saw that either the parent's and child's goals were aligned (e.g., "Cami's dad wants her to put her clothes away, and Cami is happy to do so") or misaligned (e.g., "Cami's father wants her to put her clothes away, and Cami really does not want to"). Children were asked to predict how the child protagonist would respond to the parent's directive among three options: compliance, loophole, or non-compliance (behaviors were not labeled as such or color-coded for participants).*

in a diagonal from the top left of the screen to the bottom right. As each behavior was described, the corresponding picture was highlighted. The behaviors were always described from top left to bottom right, but which behavior was in which location on the screen was randomized across trials. (See Fig. 7.)

Children indicated their prediction by clicking on the corresponding picture themselves, or pointing and having their parent click. Children were familiarized with this response method beforehand during a warm-up phase where they predicted an animal character's snack and hiding spot by clicking on or pointing to one of three on-screen options.

In the test phase, children were presented with three stories in which the parent's and child's goals were aligned, and three stories in which goals were misaligned for a total of six trials. Order was quasi-randomized across participants with the constraint that the first and second trials were always from opposite conditions – e.g., if a child saw the Aligned condition first, the second trial would be Misaligned and vice versa.

### Results of Study 3, Prediction of Loopholes



*Figure 8.* Results of Study 3: Loophole Prediction. (A) Children's mean predictions collapsed across age for each behavior (green is compliance, orange is loophole, and red is non-compliance) by goal alignment condition. Error bars are 95% bootstrapped CIs of the subject means (i.e., means per subject per condition as each subject provides 3 responses per condition). (B) Children's mean predictions for each behavior by goal alignment and age in months. Lines are fit model lines for each behavior (again green is compliance, orange is loophole, and red is non-compliance) using the R-package `ggplot::geom_smooth` using method "loess". Points are mean predictions for each behavior binned by age in years (9.5-year-olds are grouped with 10-year-olds to give a better sense of how the data extends to the end of the age range).

## Results and Discussion

As a reminder, when goals were aligned, we expected that children would be most likely to predict compliance (over both loopholes and non-compliance), and that they would be more likely to predict compliance when goals were aligned than misaligned. When goals were misaligned, we were uncertain ahead of time which behavior would be preferred overall, but we anticipated that children would be more likely to predict loopholes and non-compliance, compared to when goals were aligned. We also predicted that if there were an effect of age, it would likely be an interaction with goal alignment, such that children's predictions of loopholes vs. compliance and non-compliance would change with age when goals were misaligned, but not when goals were aligned.

As can be seen in Fig. 8(A), collapsing across age, children overwhelmingly predicted

compliance over both loopholes and non-compliance in the Aligned condition (compliance: 90.6%, 95% CIs: [86.4%, 93.5%]; loopholes: 6.0%, 95% CIs: [3.6%, 8.8%]; non-compliance: 3.4%, 95% CIs: [1.4%, 5.8%]). In the Misaligned condition, compliance was still preferred overall, but it was predicted at a much lower rate, while loopholes and non-compliance were predicted at much higher rates (compliance: 46.1%; 95% CIs: [39.0%, 53.4%], loopholes: 26.9%, 95% CIs: [21.3%, 33.1%]; non-compliance: 26.9%, 95% CIs: [20.5%, 33.1%]).

To investigate the effects of age on children's predictions, we conducted a pre-registered, confirmatory Bayesian multinomial (categorical) regression predicting participants' action choice (categorical, 3-levels: compliance, loophole, and non-compliance, with loophole dummy-coded as reference) from fixed effects of goal alignment (2-level factor: aligned and misaligned, with aligned dummy-coded as reference), age in months (continuous and centered), and their interaction, with maximal random effects (i.e., random intercepts and effects of goal alignment by subject and random intercepts, and effects of goal alignment, age, and their interaction by story/scenario). We also fit a simpler additive model, and conducted a formal model comparison, which slightly preferred the interactive model (the posterior model probability was 52% for the interactive model), so we report the results from this analysis (i.e.,  $action \sim goal\_alignment * age\_centered + (1 + goal\_alignment || subject) + (1 + goal\_alignment * age\_centered || story)$ ).

This analysis confirmed that in the Aligned condition, children were more likely to predict compliance than loopholes ( $\beta = 4.44$ , 95% CIs: [3.44, 5.66]) and showed that this tendency did not change with age ( $\beta = 0.03$ , 95% CIs: [-0.02, 0.08]). This analysis also confirmed that children's tendency to predict compliance over loopholes decreased in the Misaligned vs. Aligned condition ( $\beta = -3.92$ , 95% CIs: [-5.38, -2.63]), and showed that this difference across conditions increased with age ( $\beta = -0.07$ , 95% CIs: [-0.14, -0.01]), meaning that as children got older they were more likely to predict loopholes (compared to compliance) in the Misaligned condition (as can be see in Fig. 8(B)). Refactoring condition so that Misaligned was the reference category and refitting the model showed that within this condition, children were no more likely to predict compliance or loopholes ( $\beta = 0.52$ ,

95% CIs:  $[-0.26, 1.26]$ , and that this did not change with age ( $\beta = -0.04$ , 95% CIs:  $[-0.08, 0.01]$ ). In short, within conditions the difference between compliance and loopholes did not change with age, but *across* conditions the difference between compliance and loopholes grew significantly larger with age (likely due to the fact that within the Aligned condition the difference got a bit bigger and within the Misaligned condition the difference grew smaller).

Now we turn to the difference in children's predictions of non-compliance and of loopholes. The model revealed that in the Aligned condition, collapsing across age, children were less likely to predict non-compliance than loopholes ( $\beta = -1.94$ , 95% CIs:  $[-3.61, -0.69]$ ) and this tendency increased with age ( $\beta = -0.08$ , 95% CIs:  $[-0.15, -0.02]$ ). Across conditions, there was no difference in children's tendency to predict non-compliance vs. loopholes ( $\beta = 1.29$ , 95% CIs:  $[-0.21, 2.99]$ ), and likewise children's tendency to predict non-compliance vs. loopholes across conditions did not change with age ( $\beta = 0.06$ , 95% CIs:  $[-0.02, 0.14]$ ). In the Misaligned condition, children were no more likely to predict non-compliance compared to loopholes ( $\beta = -0.53$ , 95% CIs:  $[-1.32, 0.16]$ ) and this tendency did not change with age ( $\beta = -0.02$ , 95% CIs:  $[-0.06, 0.02]$ ).

The analysis described assumes a linear effect of age, but as can be seen in Fig. 8(B), the change in relative preference for loopholes vs. compliance and loopholes vs. non-compliance in the Misaligned condition, is not a simple increase or decrease. Instead, it appears more parabolic than linear. While children initially predicted compliance more often than loopholes, their prediction of loophole behavior increased from age five to about eight. By age seven and eight, children seemed to predict loopholes more than non-compliance and at similar rates to compliance, but then the frequency of predicting loopholes began to decrease. By age nine and ten, children appeared to predict all three behaviors at similar rates, with compliance numerically preferred.

To investigate this apparent non-linearity in children's predictions of loopholes in the Misaligned condition, we ran an exploratory analysis to test for a parabolic effect of age. We filtered the data to only consider the Misaligned condition, and recoded children's pre-

dictions so that they were binary (loophole or not loophole). We then fit a Bayesian logistic regression predicting the rate of children's loophole predictions (coded as '1' for loophole and '0' for compliance and non-compliance) from fixed effects age in months (centered and scaled) and age in months (centered and scaled) squared. This analysis found that age alone was not significant ( $\beta = 0.49$ , 95% CIs:  $[-0.01, 1.03]$ ), but that the effect of age squared was significant and negative ( $\beta = -0.75$ ,  $[-1.36, -0.20]$ ), indicating a convex parabola. This analysis, though exploratory, suggests that children's tendency to predict loopholes in the Misaligned condition did indeed significantly increase, and then significantly decrease with age.

In summary, when goals were aligned, children predicted that another child would be more likely to comply with their parent's directive (rather than not comply or exploit a loophole). When goals were misaligned, collapsing across age, children were still most likely to predict compliance, but their tendency to predict both loopholes and non-compliance greatly increased compared to when goals were aligned, in line with our experimental predictions. We also expected there to be an interaction with age such that children's tendency to predict loopholes compared to compliance and non-compliance would be more likely to change with age in the Misaligned condition than in the Aligned condition. The effect of age was indeed significantly different across conditions, and an exploratory analysis suggested that the effect of age on children's tendency to predict loopholes is parabolic, such that across the age range studied, the frequency of loophole prediction first increases and then decreases. So while compliance was overall preferred when age is collapsed, looking across age this was not consistently the case: children started out preferring compliance, then in the middle of the age range studied, they were equally likely to predict loopholes or compliance, and then their preference for loopholes dropped off. This developmental pattern in children's predictions of loopholes is consistent both with the pattern reported in Study 1 (i.e., onset at five to six, peak at seven to eight, tapering off at nine to ten) and the observed pattern in Study 2 (children around age four think loopholes will receive similar punishment to non-compliance and with age predict that it will receive less, with suggestive

evidence though that this difference might expand and then contract at the upper end of the age range around nine or ten).

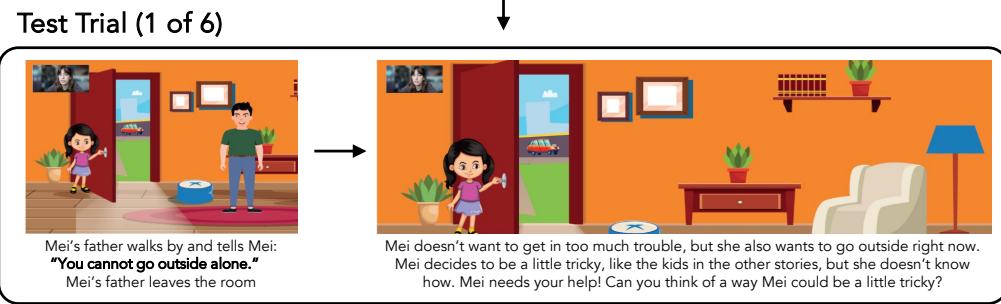
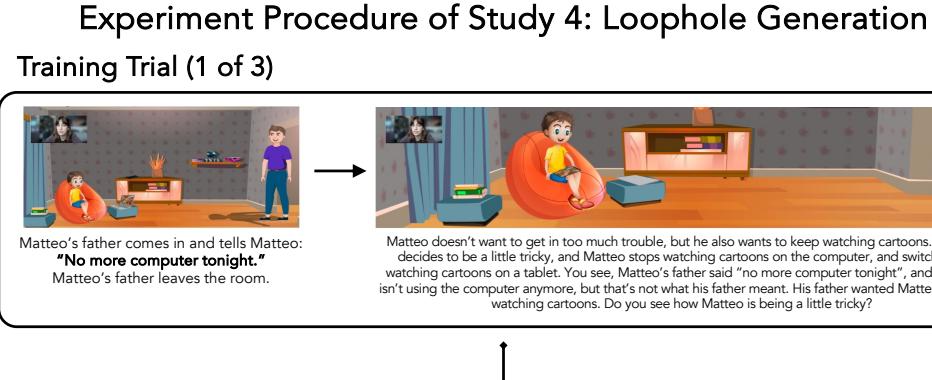
Studies 2 and 3 assess children's understanding of loopholes more directly than the parent survey, and while they line up with its results, they do not assess children's actual production of loopholes. Both of these studies give children the loophole behavior, rather than asking them to come up with it themselves. In the next study, we directly probe children's abilities to come up with loopholes.

#### **Study 4: Children's own generation of loopholes**

In Study 4, we examined children's abilities to generate loopholes. We explained to children ages five up until ten what we mean by loophole behavior (referring to it as "being a little bit tricky or sneaky"). We presented children with stories in which a child protagonist does not want to follow the directive of a parent but also does not want to get into too much trouble, and then asked the child participants to come up with ways of helping the protagonist figure out how to be a little bit tricky or sneaky with their parent (i.e., find a loophole). We predicted that children would increasingly be able to come up with a relevant loophole as they age.

### **Methods**

**Participants.** We recruited 60 5- to 9-year-olds ( $M_{age}$ : 7.60 yrs, range: 5.09 to 10.02 yrs; 45% female; other demographic data not collected) through the Children Helping Science platform. Children participated in a synchronous, researcher-moderated experiment online over Zoom from December 2021 to April 2022. An additional 15 participants were recruited but excluded from analysis due to experimenter error ( $N = 2$ ), parental interference ( $N = 1$ ), and prior participation in a similar study, which was difficult to verify beforehand ( $N = 12$ ). Additional 5-year-olds ( $N = 7$ ) participated but were excluded from analysis, because they were accidentally permitted to sign up after we had already recruited the pre-registered quota of 5-year-olds. We did not include these children so as not to exceed



*Figure 9. Structure of Study 4, Loophole Generation. Participants were introduced to stories in which a parent gave a directive to their child. Participants were told that the child protagonist did not want to comply with their parent's request but also didn't want to get into too much trouble, and so wanted to be a little tricky or sneaky. Children listened to three training trials in which an experimenter showed participants how a child in each story could be a little tricky (by exploiting a loophole, though the word 'loophole' was never used). Children then completed six test trials in which the child protagonist wanted to be a little tricky but didn't know how, and the child participants were asked to help the protagonist find a way to be a little tricky.*

our pre-registered sample size. The sample size, inclusion and exclusion criteria, as well as the hypotheses and analyses were all preregistered on the Open Science Framework.

**Procedure.** We again used the same twelve scenarios used in Studies 2 and 3, in which a parent issues a directive to a child. Participants were presented with nine of these scenarios as stories in a keynote presentation that the experimenter shared over Zoom and narrated. Three stories were example trials, and six were test trials. Which nine of the twelve scenarios children saw, as well as which of those nine served as the three example stories and the six test stories were randomized across participants. The parent directives were pre-recorded, using the same recordings as were used in Studies 2 and 3,

**Results of Study 4, Generation of Loopholes  
(Examples of Loophole Responses)**

"It's time to get in the tub."	"Dinner is soon, so do not eat all of the popcorn."	"No more m&m's today."
 <p><i>How can Bo be a little tricky?</i></p> <p>"Hmm he gets in the tub and plays with legos without filling it up." (6-year old)</p> <p>"Um he could put like his feet in the tub and still play legos." (9-year old)</p> <p>"Cause he could take his legos with him and then he could just sit in the tub and not turn on the water." (7-year old)</p> <p>"He can um- he can get in the tub and play with his legos without actually bathing or taking off his clothes." (8-year old)</p>	 <p><i>How can Sierra be a little tricky?</i></p> <p>"She could eat most of the popcorn, but not all of it." (10-year old)</p> <p>"Eat all of the popcorn except for one piece." (9-year old)</p> <p>"So she eats not all of the popcorn but some of it. Um almost all of it, but just one bite left." (5-year old)</p> <p>"She could leave one piece of popcorn in the bowl." (9-year old)</p>	 <p><i>How can Gemma be a little tricky?</i></p> <p>"Go get some skittles." (7-year old)</p> <p>"She could eat a different sweet that isn't m&amp;m's." (5-year old)</p> <p>"Because m&amp;ms is like chocolate, then she could take chocolate and she could eat chocolate." (7-year old)</p> <p>"There's gummy bears on the table, and maybe she can eat those because those are sweets and she's not eating the m&amp;ms." (9-year old)</p>

Figure 10. Study 4: Loophole Generation Results (Examples of Loophole Responses). Participants were asked to help a child protagonist be “a little tricky or sneaky” when the child protagonist did not want to comply with a parent’s request. Here, we show randomly selected examples of children’s verbal responses coded as loopholes, for 3 of the 12 trials.

so the experimenter did not say them and they were consistent in tone and pitch across participants.

The study began with a warm-up where children were asked to help the experimenter figure out what an animal character would do from several options, giving children practice generating behavior for a character in a story and saying it out-loud. The experimenter then explained that they would hear stories where children needed help being ‘a little tricky or sneaky’ with their parents. We used the phrasing ‘tricky or sneaky’ as some children in Study 2 used this language to describe a child who exploited a loophole. The word “loophole” was never explicitly used in the study.

In the three example trials, children were shown stories of children exploiting loopholes to demonstrate what was meant by ‘being a little tricky or sneaky’. For each story, after the parent in the story gave the directive, children were told that the child protagonist didn’t want to get into too much trouble, but also did not want to do as they were directed, and

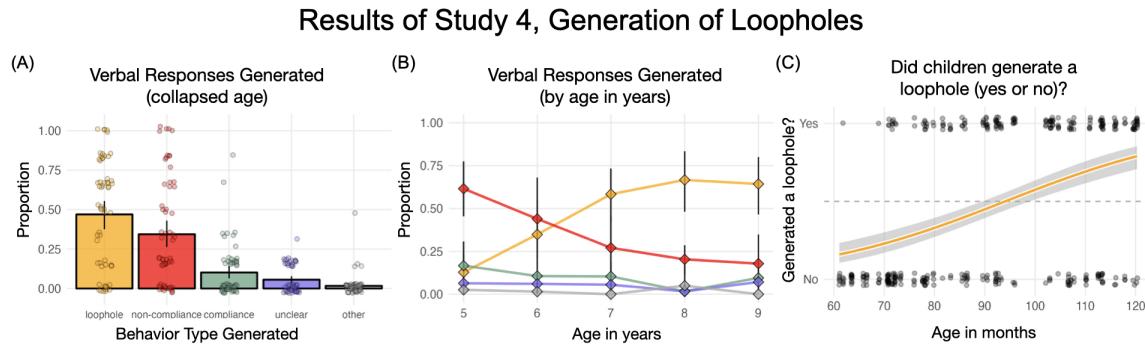
so they decided to be ‘a little tricky’. The experimenter then showed the corresponding loophole behavior for that story, and in the first two example trials, explained how the child was being ‘a little tricky’ by technically doing what the parent said, but not what they intended. For the third example trial, participants were asked to explain how the child in the story was being a little tricky and then, regardless of whether or not they responded, the experimenter provided an explanation.

In the six test trials, children were not shown the loophole behavior and instead were tasked with helping the child protagonist figure out how to be ‘a little tricky’ (i.e., generating a loophole). For each story, children were told that the child protagonist did not know how to be a little tricky, and child participants were asked if they could think of way the child protagonist could be little tricky (see Fig. 9).

## Results and Discussion

We predicted that overall children would be able to spontaneously generate loophole behavior for a given directive, and that this ability would increase with age. To test these hypotheses, we first did a pre-registered exploratory classification of children’s responses into one of five categories, based on whether the response (1) complied with the parent’s request (*compliance*), (2) did not comply with the parent’s request (*non-compliance*), (3) exploited a loophole in the parent’s request (*loophole*), (4) was unclear and could not be coded as one of the other three behaviors (*unclear*), or (5) was non-sensical or irrelevant (*other*) (inter-rater agreement between the two first authors was  $\kappa = 95\%$  and  $\kappa = 83\%$  between a first author and a blind-coder, see Supplementary Material for coding details).

As can be seen in Fig. 11(A), collapsing across age, children were able to generate loopholes on a substantial portion of trials (47%, 95% CIs: [38%, 55%],) and reliably more often than non-compliance (36%, 95% CIs: [36%, 43%]). Looking at these results by age (Fig. 11(B)), we observe a clear developmental trend, such that children at age five are producing non-compliance over loopholes (62% vs. 13%), but this switches by age seven (27% non-compliance vs. 58% loopholes) and continues to increase.



*Figure 11.* Study 4: Loophole Generation Results. (A) The proportion of trials in which children generated responses that were considered examples of "loopholes" (orange), "non-compliance" (red), "compliance" (green), "unclear" (purple), or "other" (gray), collapsing across age. Error bars are 95% bootstrapped CIs of the subject means (i.e., each participant provided up to six responses, so we first calculated the mean for each behavior category by subject and then bootstrapped those subject means by category). (B) The proportion of trials for each behavior category by age in years. Error bars are 95% bootstrapped CIs of the mean. (C) Children's generation of loophole responses by age in months with a binary classification of the response as 'yes, a loophole' or 'no, not a loophole'. Dots are individual subjects' responses (up to six responses per subject); lines are fit model lines using the R-package `ggplot::geom_smooth` using method "glm".

To test the effect of age on loophole production, we re-coded children's responses as either 'yes, a loophole' or 'no, not a loophole' (i.e., all other behaviors) and conducted a pre-registered confirmatory Bayesian mixed-effects logistic regression predicting children's responses (0 or 1) from a fixed effect of age (continuous and centered) with a maximal random effects structure (random intercept by subject and random intercept and effect of age by story). This analysis confirmed that children's ability to generate loopholes significantly increased with age ( $\beta = 0.07$ , 95% CI: [0.04, 0.11]).

In examining children's 'non-compliance' responses, we noted that they tended to fall into two distinct types, suggesting that 'non-compliance' may be too crass a category. The types of non-compliance observed were (1) direct non-compliance and (2) a category we refer to as 'sneaky non-compliance'. To see the difference between these, consider a child who was told they have to eat some peas before having more pizza, and contrast a child who simply refuses and continues eating pizza with a child who waits for their parent to leave the room, throws the peas in the garbage bin, and then returns quickly to their seat

with a plate free of peas. Both children refused to fulfill the request and would be classified as ‘non-compliance’ under our original coding scheme, but the second child is engaging in what we term ‘sneaky non-compliance’ (not fulfilling the request, but in a way that makes the parent unaware that this has happened). When using a more fine-grained analysis, we find that the vast majority of ‘non-compliance’ behaviors generated by children at age five are actually ‘sneaky non-compliance’, and it is ‘sneaky non-compliance’ that decays by age six, while direct non-compliance stays roughly the same throughout all ages examined (see the Supplemental Materials for more details on this analysis). This exploratory analysis suggests that the youngest children in the age range studied understood the assignment of being “sneaky or tricky” but did not grasp the particular intended notion of “sneakiness or trickiness”, that is following the letter but not the spirit of the directive.

In summary, children in Study 4 were able to actively generate loopholes on the spot given a directive (see Fig. 10 for examples). This ability showed a clear and significant increase with age: children generated non-compliance responses more often than loopholes at the age of five, and then loophole production steadily increased, such that by age eight this pattern has switched and they reliably generated loopholes more often than non-compliance.

This developmental trajectory of children’s ability to generate loopholes is again largely consistent with parental experiences of their children’s loophole behavior gathered in Study 1, as well as children’s evaluations of and tendency to predict loopholes in Studies 2 and 3. All of these studies point to the idea that children’s understanding and ability to use loopholes emerges around five to six years of age and that seven and eight years of age is a period of particularly frequent use (at least with their parents).

## General Discussion

Intentional misunderstandings or loopholes allow people to get around situations where they neither want to comply nor want to incur the costs of not complying with another person’s directive (Bridgers et al., 2023; Qian et al., 2024). This function of loopholes is especially important in situations where there is a power imbalance, as in many

of children's relationships with the adults around them. Thus, we set out to examine the development of loophole behavior as a window into how children's developing social and cognitive capacities enable them to navigate the gray area between compliance and defiance and handle the challenges of cooperation.

Across four studies, we established that loopholes are an ecologically distinct behavior in childhood, different from both compliance and non-compliance, and we mapped the developmental trajectory of loophole behavior from age four to ten years. In Study 1, parents reported that their children made use of loopholes across a variety of utterances and situations, beginning around age five to six years, peaking around seven to eight, and tapering off around nine to ten (loophole-use persists into adulthood, but at least with one's parents, this use appears to decline into adolescence). Directly probing children's abilities revealed that their capacity to generate valid loopholes improves from ages five to ten (Study 4). We also examined *when* and *why* children engage in loopholes, finding that overall children's reasoning converged with that of adults' (Bridgers et al., 2023). Children expected loopholes to mitigate social costs compared to non-compliance (with this distinction most clearly seen from five to eight; Study 2), and they predicted that loopholes would be used selectively when social partners had conflicting goals (with this tendency appearing to increase from five to eight and decrease from eight to ten; Study 3). Altogether, these studies reveal that children learn to distinguish loopholes from compliance and non-compliance, that children broadly exploit loopholes in their daily lives, and that children's loophole production and comprehension may be tightly linked, with five to seven years being a critical time for emergence and development.

The four studies provide converging evidence of how children's understanding and use of loopholes changes from early to middle childhood. Parents' self-reports on the emergence and frequency of loopholes in Study 1 align with children's own understanding and generation of loopholes in Studies 2 and 4. In Study 2, 4-year-olds did not appear to differentiate loopholes from non-compliance in terms of trouble, while 5- to 8-year-olds expected that loopholes would incur less trouble, aligning with parent reported age of onset (5.6 years) and

peak frequency (7.4 years). In Study 4, children's abilities to come up with loopholes on the spot also steadily increased from age five to seven, such that by age eight, children reliably generated loopholes more often than non-compliance. Together these findings suggest that children's ability to understand loopholes and their ability to produce loopholes emerge in tandem. They also build upon and expand existing research about the opposite of loophole behavior (Bregant et al., 2019), and indicate that children consider behavior that violates *either* the letter or the spirit of the law as deserving of more leniency than non-compliant behavior that violates both.

While the *understanding* of loopholes emerges around age five and increases in the years to follow, the *use* of loopholes appears to have a rise-and-fall pattern (both in children's prediction of their use and parents' report of their actual use). In Study 3, children's predictions of loopholes when goals were misaligned significantly increased from five to about eight years of age, but then significantly decreased from about eight to ten, aligning with parent reported age of peak frequency (7.4 years) and of offset (9.3 years). This decline in rates of prediction lends credence to the suggestive evidence from Study 2 that 9-year-olds, like 4-year-olds, may not differentiate loopholes from non-compliance in terms of trouble. If both 4- and 9-year-olds do not expect loopholes to mitigate punishment, it is likely for different reasons as our work suggests that 4-year-olds may not understand the difference between loopholes and non-compliance, while 9-year-olds certainly do. More work is needed to establish the reliability and robustness of these patterns, and to more directly test *why* older children possibly go back to believing loopholes will be as costly as non-compliance.

The correspondence between parent self-report and children's actual behavior in our experiments speaks to the value of combining these methods. While parent surveys provide an indirect measure of children's behavior, and thus cannot be relied on as the sole source of evidence, they can help to establish the ecological validity of behavioral phenomena and ground hypotheses that can be tested experimentally. Moreover, even when parent reports diverge from children's behavior in experiments, these reports still provide insight

into parents' beliefs about their children. Parent beliefs are a valuable and informative data source unto themselves, as they may play a causal role in shaping children's cognition and behavior (e.g., see Haimovitz & Dweck, 2016).

Loopholes operate in the liminal space between compliance and defiance. But they are not alone in this space and share it with other behaviors used to get around social misalignment. Unlike deception (such as the 'sneaky non-compliance' that children came up with in Study 4), these behaviors can mitigate punishment even if a social partner is aware of them. As just a few examples of behaviors already studied in other work, excuses and apologies can also decrease punishment (Schlenker & Weigold, 1992; Snyder & Higgins, 1988), as does partial compliance (Bridgers et al., 2023). While loopholes are part of a broader picture of behavior, they are still a particularly interesting phenomenon within this picture, because their use and understanding relies on a coherent, complex set of social and cognitive processes that are particularly relevant for development.

Engaging with loopholes requires that children integrate an array of social and cognitive processes that include (but are not limited to): an understanding of pragmatics, reasoning about utilities and beliefs, and trading-off utilities for joint planning. Given this, we expected that the ability to recognize and exploit loopholes would dovetail with developmental trajectories in these related abilities. We found that children begin to understand and distinguish loophole behavior from non-compliance between the ages of five and seven. This is the age range within which we roughly expected the ability to emerge, given that past work suggests that children's ability to generate relevant alternative utterances and interpretations for a given utterance is improving during this time, enabling them to better draw pragmatic implicatures (Barner et al., 2011; Bohn & Frank, 2019) and comprehend puns, irony, and metaphor (McGhee, 1974; Shultz & Horibe, 1974; Winner et al., 1988). This age range is also when children are developing a more sophisticated, higher-order Theory-of-Mind: By age five children understand false beliefs, but from five to seven they begin to explicitly represent other people's beliefs about other people's beliefs (e.g., Tomasello, 2018), which may be critical for estimating the probability that one's intentional

misunderstanding may be perceived as genuine confusion.

While our results are generally in line with the expectations set by prior literature, we wish to consider additional, less obvious possibilities for why the ages of five-to-seven may be a particularly important time for the emerging understanding and use of loopholes. We note that the following is speculative and raised for the purposes of considering new lines of research. Specifically, consider that the emergence of loopholes overlaps with the age where children begin to use and understand some forms of counterfactuals (Beck & Crilly, 2009; Rafetseder, Cristi-Vargas, & Perner, 2010) (though the specific developmental timelines for counterfactual reasoning are under debate, see e.g., Rafetseder, Schwitalla, & Perner, 2013), and where they begin to reason explicitly about different modal possibilities (Leahy & Carey, 2020; Leahy & Zalnieriunas, 2021). One over-arching possibility tying loopholes, counterfactuals, modals, and puns together is the growing development of executive function, and specifically working memory. For the understanding of many counterfactuals, children need to keep in mind (at least) a representation of a current world, as well as an altered world. For comparing different modal possibilities, children need to keep in mind (at least) two different possible actions with different probabilities of success. For puns, irony, and metaphor, children need to keep in mind (at least) a representation of a literal interpretation, as well as an intended interpretation. And more pertinent to our purposes here, for loopholes children need to keep in mind (at least) the actual intended request, as well as the supposed misinterpreted request (we emphasize the ‘supposed’ here: if children truly did misunderstand a request, they would then simply be confused, and also would only need to hold one thing in mind - what they believe the intended request to be). The idea that the emergence of loopholes relies on an increase in working memory is intriguing, but remains to be tested, and even for counterfactual reasoning there is no acceptance of the idea that working memory differences in middle childhood are a driving factor. In our ongoing research, we are investigating more directly the relation between loophole behavior and other cognitive abilities related to the need of keeping multiple options in mind. Specifically, we are seeing if there is a correlation between counterfactual and false belief

reasoning and children's tendency to evaluate loopholes as less costly than non-compliance.

One aspect of loopholes that can get lost in dry discussions about ‘multiple interpretations’ and ‘goal misalignment’ is that they are *funny*. This is a reliable, distinguishing feature of loopholes: Both children and adults consider loopholes amusing, as opposed to compliance and non-compliance (Bridgers et al., 2023, 2021). While our work shows this empirically, it does not answer *why* loopholes are funny. This is not for lack of possible answers, but due to an abundance of them. Some theories of humor suggest that humor is based on violations of schemata (Deckers & Buttram, 1990), an element that loopholes capitalize on as they make use of an unexpected linguistic interpretation. Other theories posit the importance of incongruity (Forabosco, 1992; Hurley, Dennett, & Adams, 2011), the presence of two incompatible ideas or meanings, which is also true of loopholes: the person is either cooperative (but confused) *or* uncooperative (and intentionally misunderstanding). McGhee et al. (1983) proposes that by age seven, children can identify the “multiple meanings” and ambiguity needed to comprehend riddles and jokes. Kao, Levy, and Goodman (2016) build on the incongruity theories, showing that distinctiveness, the degree to which you can attribute each of the incompatible meanings to different parts of the context, is also needed to explain why puns are funny (e.g., in the pun, “The magician got so mad he pulled his hare out.”, ‘magician’ supports ‘hare’ but ‘mad’ supports ‘hair’). Qian et al. (2024) extend this model of puns to loopholes, suggesting that loopholes may be humorous because unlike compliance or non-compliance, they reveal that there is potential ambiguity in the person’s intent by highlighting that the directive has different possible interpretations each supporting a different intent (unintended supporting cooperative, and intended supporting uncooperative). Still, all this talk about ‘conflicting interpretations’ leaves open the question of *why* such things are humorous. Here, we would suggest (in line with other theories, Darwin, 1872; Hurley et al., 2011) that the primary purpose of humor in these situations is communicative, sending a message to another party that one recognizes that multiple interpretations are possible, but the interpretation upon which they are acting, they only hold in a pretend sense (as they do indeed understand what was really

meant). More work is needed to see if this proposal also captures what children find funny about intentional misunderstandings.

We found that children expect loopholes to result in less cost than non-compliance (though this expectation may diminish at the oldest ages studied). Similar to humor, this empirical finding is in line with adults' understanding of loopholes (Bridgers et al., 2023; Qian et al., 2024), but again this observation does not answer *why* loopholes lead to reduced punishments. One possible account is that children are engaging in "plausible deniability": they believe there is a chance that their intentional misunderstanding will be seen as a genuine misunderstanding, and they know that genuine lack of intent leads to reduced punishment (Cushman, 2008). In other words, according to this explanation, the child thinks, "If I use this behavior, my parent *might* suspect I am genuinely confused, and you don't get punished for being genuinely confused," and the parent, for their part, upon observing the loophole, is indeed unsure of the child's intent, and so is more lenient. We note that previous work has already deeply explored indirect speech and the use of plausible deniability, mostly focusing on the person *crafting* the original message who can claim that they did not intend a particular meaning (e.g., Pinker, Nowak, & Lee, 2008), whereas here the focus would be on children engaging in this same process as the person *receiving* the message who can claim that they did not understand a particular meaning.

While plausible deniability is a plausible explanation for why loopholes get children out of trouble, another possibility is that children are engaging in "implausible deniability", similar to the account proposed in Qian et al. (2024) for loophole use in adults. We refer the interested reader to Qian et al. (2024) for the details of a computational model of implausible deniability and how it differs from some models of plausible deniability, but briefly here: In plausible deniability, the person on the receiving end of an ambiguous behavior (e.g., "Nice store you have here, shame if it burns down") is somewhat unsure if they are on the receiving end of a negative intention (a threat to burn down their store) or a positive intention (a compliment), or they are unsure if a reasonable third party would see the indirect action as negative (i.e., "I know this is a threat, but would other people know?", see e.g., Bonalumi,

Bumin, Scott-Phillips, & Heintz, 2023; Hall & Mazzarella, 2023; Pinker et al., 2008). In implausible deniability, however, no one in the interaction is really in doubt about the negative intention of an indirect behavior nor about how other reasonable people would see it (i.e., “I know that you’re trying to threaten me, and so would anyone else”). But, all parties, nevertheless, expect punishment to be tied to the interpretation of a naive, literal observer. Such a naive or literal observer is similar to a literal listener at the bottom floor of a recursive reasoning process (e.g., the Rational Speech Act approach to communication, see Goodman & Frank, 2016b). One may then well ask why punishment is tied to the standards and inferences of such a non-existent, naive person. As expanded upon in Qian et al. (2024) and discussed in Ullman-Margalit (1983), we note that this is similar to how in legal frameworks one imagines a person with higher standards of conviction to assess blame and punishment (e.g., ‘beyond a reasonable doubt’).

Let us consider again the child watching movies on their tablet, who hears ‘Time to put the tablet down’, and puts the tablet down on the table – only to continue watching movies. If the child is engaging in plausible deniability, they would expect that their parent *might* believe that they truly did not understand what was said, and since confused children don’t deserve punishment, the loophole would lead to less trouble than non-compliance. By contrast, if the child is engaging in implausible deniability, they would expect their parents to know that they were not confused, but since a naive, literal observer *might* believe the child genuinely misunderstood the command, and since punishment is tied to the standards of this observer, the loophole would get them into less trouble. Further developing the Qian et al. (2024) model to incorporate the development of language and pragmatics (e.g., Bohn, Tessler, Kordt, Hausmann, & Frank, 2023; Bohn, Tessler, Merrick, & Frank, 2021) and empirically testing if children’s engagement with loopholes follows the assumptions of implausible deniability could shed light on children’s changing use of loopholes and the implausible v. plausible distinction.

Still another possibility for why loopholes reduce punishment is the fact that they are funny. As discussed, the humor elicited by using a loophole likely signals that all involved

understand that a genuine misunderstanding is not the case. If so, this would reduce even further the notion of *plausible* deniability, since slyly smiling or laughing when using a loophole would be a tip-off that one is not actually confused. But, such humor in itself may reduce social penalties: In Study 1, parents noted that when their children came up with loopholes, they laughed or chuckled in response, and mentioned the “added brain power”, or the cleverness and cunning needed to exploit loopholes. The proposed pathway is then along the following lines: children exploiting a loophole also communicate humor via affect to signal they’re holding a misinterpretation in pretense. This shared humor with the person making the request reduces potential costs due to positive affect, and/or because finding a loophole is seen as creative and clever (like finding a pun), leading to reduced costs.

But if loopholes are funny, and children can (plausibly or implausibly) claim that they misunderstood the situation, then why do older children (9- and 10-year-olds) stop using them and stop believing they can get you out of trouble? In our studies, parents reported that their children stopped using loopholes around age nine, children’s tendency to predict others’ loophole use declined from eight to ten, and by age ten, children no longer expected loopholes would incur less trouble than non-compliance. These observed declines don’t seem to line up with the observations that adults continue to use loopholes, predict others will use loopholes, and expect loopholes to reduce social costs (Bridgers et al., 2023; Qian et al., 2024). One possible reason for this difference is that here we focused exclusively on the use of loopholes in parent-child interactions, while prior work with adults explored a diverse range of relationships (e.g., both egalitarian and hierarchical relationships across familial, professional, housing, and collegial settings). It might be specifically within parent-child dynamics that by age ten, children no longer believe loopholes will incur less punishment than outright disobedience. This shift in expectation could stem from real-world experience: older children’s attempts at loophole exploitation may simply exasperate parents, leading to harsher consequences. Such exasperation may be driven by a number of overlapping factors: perhaps it is the sheer frequency of loopholes that is working to make them less palatable, or perhaps as children get older the standards for cleverness and humor increase, such that

what was previously a marker of wit no longer meets the bar. Or possibly when children are younger, there is greater plausible deniability such that loopholes may sometimes be viewed as genuine misunderstandings; any such plausible deniability would presumably diminish over time. These possibilities could be disentangled with further investigation, such as by surveying older children, teens, and their parents about loopholes vs. non-compliance, and by exploring children's loophole use across a broader space of relationships varying in hierarchy (e.g., older vs. younger siblings) and intimacy (e.g., parents vs. babysitters).

Our findings advance the understanding of the development of loopholes, but they have several limitations which form the basis for future work. Some of these limitations we have just discussed: we only focused on one type of relationship and power-dynamic, and we didn't investigate loophole use beyond age ten. Though the parent-child relationship is a major relationship in development and our samples covered a fairly large age range, studying children's loophole use across more relationships and into adolescence would paint a fuller picture of the development of loophole behavior from early childhood to adulthood.

Another major limitation of our work is that while we sought to obtain a diverse sample representative of a broad population, the populations we worked with were based in the United States. It is likely that beyond the development of the social and cognitive processes we already mentioned, differences in culture and parenting styles will play a role in the emergence, understanding, and use of loopholes. In particular, prior work has shown that country of residence and ethnicity account for substantial differences in the types of discipline parents use (Lansford, 2022; Silveira, Shafer, Dufur, & Roberson, 2021). Consequently, differences in parenting styles and disciplinary practices may influence children's ratings of trouble and behavioral predictions (Studies 2 and 3). For example, while we found that children overall predict compliance both in cases of goal alignment and goal misalignment, the particular frequency of expected compliance may depend on the emphasis a particular culture places on obedience. Generally, we expect differences in culture and parenting style to have bigger roles in the evaluation and use of loopholes, rather than the ability to understand or come up with a loophole in the first place, but future work is needed

to examine and address this hypothesis.

In the philosophical paradox known as ‘Buridan’s Ass’, an equally hungry and thirsty donkey is placed exactly between hay and water. Being unable to choose which way to go, the donkey eventually expires. Children often find themselves with two unappealing options, between submitting and refusing. While younger children may be forced to simply pick one of the options, as they get older children become much smarter asses.

### Acknowledgments

The authors wish to thank Sienna Radifera, Sofia Riskin, and Elizabeth Choi for their help coding and recording stimuli. We would also like to thank members of the MIT Early Childhood Cognition Lab and of the Harvard Computation, Cognition, and Development Lab for useful discussions, and MH Tessler and Natalia Vélez for helpful feedback on the manuscript. We are grateful to Children Helping Science and the children and families who participated in this research. This research was funded by a MIT Simons Center for the Social Brain Postdoctoral Fellowship (SB), the Jacobs Foundation (TDU), and a NSF Science of Learning and Augmented Intelligence Grant 2118103 (TDU, LS).

The data and analytic code necessary to reproduce the analyses presented here are not publicly accessible. Data are available from the first authors upon reasonable request. The data is available on the Open Science Framework at the following URL: [https://osf.io/rwgmx/?view\\_only=51d508a0a4684506932d1f729490f096](https://osf.io/rwgmx/?view_only=51d508a0a4684506932d1f729490f096). The materials necessary to attempt to replicate the findings presented here are publicly accessible at the following URL: [https://osf.io/rwgmx/?view\\_only=51d508a0a4684506932d1f729490f096](https://osf.io/rwgmx/?view_only=51d508a0a4684506932d1f729490f096). The analyses presented here for Study 1 were not preregistered. The analyses presented here for Studies 2 to 4 were preregistered and can be accessed at the following URLs: Study 2 ([https://osf.io/du4vp/?view\\_only=d0e3e8f6e3d7435a850b3230915f2c15](https://osf.io/du4vp/?view_only=d0e3e8f6e3d7435a850b3230915f2c15)), Study 3 ([https://osf.io/p89es/?view\\_only=e38dd168cd954e3ba39cddd422c1b42a](https://osf.io/p89es/?view_only=e38dd168cd954e3ba39cddd422c1b42a)), and Study 4 ([https://osf.io/wzajy/?view\\_only=71106674e6824eef852f0304db2b6c21](https://osf.io/wzajy/?view_only=71106674e6824eef852f0304db2b6c21)).

## References

- Barner, D., Brooks, N., & Bale, A. (2011). Accessing the unsaid: The role of scalar alternatives in children's pragmatic inference. *Cognition*, 118, 84–93.
- Bates, E. (1976). *Language and context: The acquisition of pragmatics*. Academic Press.
- Beck, S. R., & Crilly, M. (2009). Is understanding regret dependent on developments in counterfactual thinking? *British Journal of Developmental Psychology*, 27, 505–510.
- Bohn, M., & Frank, M. C. (2019). The pervasive role of pragmatics in early language. *Annual Review of Developmental Psychology*, 1, 223–249.
- Bohn, M., Tessler, M. H., Kordt, C., Hausmann, T., & Frank, M. C. (2023). An individual differences perspective on pragmatic abilities in the preschool years. *Developmental Science*, 26, e13401.
- Bohn, M., Tessler, M. H., Merrick, M., & Frank, M. C. (2021). How young children integrate information sources to infer the meaning of words. *Nature Human Behaviour*, 5, 1046–1054.
- Bonalumi, F., Bumin, F. B., Scott-Phillips, T., & Heintz, C. (2023). Communication and deniability: Moral and epistemic reactions to denials. *Frontiers in Psychology*, 13, 1073213.
- Boyd, R., & Richerson, P. J. (2005). The origin and evolution of cultures. *Oxford University Press*.
- Boyd, R., & Richerson, P. J. (2009). Culture and the evolution of human cooperation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1533), 3281–3288.
- Bratman, M. E. (1992). Shared cooperative activity. *The philosophical review*, 101, 327–341.
- Bregant, J., Wellbery, I., & Shaw, A. (2019). Crime but not punishment? children are more lenient toward rule-breaking when the “spirit of the law” is unbroken. *Journal of experimental child psychology*, 178, 266–282.
- Bridgers, S., Jara-Ettinger, J., & Gweon, H. (2020). Young children consider the expected utility of others' learning to decide what to teach. *Nature Human Behaviour*, 4, 144–152.
- Bridgers, S., Qian, P., Taliaferro, M., Parece, K., Schulz, L., & Ullman, T. D. (2023, Aug). Loopholes: A window into value alignment and the communication of meaning. *PsyArXiv*. Retrieved from [psyarxiv.com/cnxzv](https://psyarxiv.com/cnxzv) doi: 10.31234/osf.io/cnxzv
- Bridgers, S., Schulz, L., & Ullman, T. D. (2021). Loopholes, a window into value alignment and the learning of meaning. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 43).
- Brownell, C., Svetlova, M., & Nichols, S. (2009). To Share or Not to Share: When Do Toddlers Respond to Another's Needs? *Infancy*, 14, 117–130.

- Buttelmann, D., Carpenter, M., & Tomasello, M. (2009). Eighteen-month-old infants show false belief understanding in an active helping paradigm. *Cognition*, 112, 337–342.
- Cortes Barragan, R., & Dweck, C. S. (2014). Rethinking natural altruism: Simple reciprocal interactions trigger children's benevolence. *Proceedings of the National Academy of Sciences*, 111, 17071–17074.
- Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, 108, 353–380.
- Darwin, C. (1872). *The Expression of the Emotions in Man and Animals* (F. Darwin, Ed.). London: John Murray.
- Deckers, L., & Buttram, R. T. (1990). Humor as a response to incongruities within or between schemata. *Humor: International Journal of Humor Research*.
- Demorest, A., Silberstein, L., Gardner, H., & Winner, E. (1983). Telling it as it isn't: Children's understanding of figurative language. *British Journal of Developmental Psychology*, 1, 121–134.
- Dunfield, K., Kuhlmeier, V. A., O'Connell, L., & Kelley, E. (2011). Examining the diversity of prosocial behavior: Helping, sharing, and comforting in infancy. *Infancy*, 16, 227–247.
- Forabosco, G. (1992). Cognitive aspects of the humor process: The concept of incongruity. *Humor - international Journal of Humor Research - HUMOR*, 5, 45-68.
- Garcia, S. M., Chen, P., & Gordon, M. T. (2014). The letter versus the spirit of the law: A lay perspective on culpability. *Judgment and Decision making*, 9, 479–490.
- Gergely, G., & Csibra, G. (2003a). Teleological reasoning in infancy: the naïve theory of rational action. *Trends in Cognitive Sciences*, 7, 287–292.
- Gergely, G., & Csibra, G. (2003b). Teleological reasoning in infancy: The naive theory of rational action. *Trends in cognitive sciences*, 7, 287–292.
- Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995, August). Taking the intentional stance at 12 months of age. *Cognition*, 56, 165–193.
- Goodman, N. D., & Frank, M. C. (2016a). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20.
- Goodman, N. D., & Frank, M. C. (2016b). Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, 20, 818–829.
- Gweon, H., Shafto, P., & Schulz, L. (2018). Development of children's sensitivity to overinformativeness in learning and teaching. *Developmental psychology*, 54, 2113.

- Haimovitz, K., & Dweck, C. S. (2016). What predicts children's fixed and growth intelligence mind-sets? not their parents' views of intelligence but their parents' views of failure. *Psychological science*, 27, 859–869.
- Hall, A., & Mazzarella, D. (2023). Pragmatic inference, levels of meaning and speaker accountability. *Journal of Pragmatics*, 205, 92–110.
- Hamlin, K. J., Ullman, T., Tenenbaum, J. B., Goodman, N., & Baker, C. (2013). The mentalistic basis of core social cognition: experiments in preverbal infants and a computational model. *Developmental science*, 16, 209–226.
- Hannikainen, I. R., Tobia, K. P., de Almeida, G. d. F., Struchiner, N., Kneer, M., Bystranowski, P., ... others (2022). Coordination and expertise foster legal textualism. *Proceedings of the National Academy of Sciences*, 119, e2206531119.
- Henrich, J. (2015). *The secret of our success: how culture is driving human evolution, domesticating our species, and making us smarter*. Princeton University Press.
- Huang, Y. T., & Snedeker, J. (2009). Online interpretation of scalar quantifiers: Insight into the semantics-pragmatics interface. *Cognitive Psychology*, 58, 376-415. doi: <https://doi.org/10.1016/j.cogpsych.2008.09.001>
- Hurley, M. M., Dennett, D. C., & Adams, R. B. (2011). *Inside jokes: Using humor to reverse-engineer the mind*. MIT press.
- Isenbergh, J. (1982). *Musings on form and substance in taxation*. HeinOnline.
- Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The Naïve Utility Calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, 20, 589–604.
- Jara-Ettinger, J., Floyd, S., Huey, H., Tenenbaum, J. B., & Schulz, L. E. (2020). Social pragmatics: Preschoolers rely on commonsense psychology to resolve referential underspecification. *Child Development*, 91, 1135–1149.
- Kao, J. T., Levy, R., & Goodman, N. D. (2016). A computational model of linguistic humor in puns. *Cognitive science*, 40, 1270–1285.
- Katz, L. (2010). A theory of loopholes. *The Journal of Legal Studies*, 39, 1–31.
- Lansford, J. E. (2022). Annual research review: Cross-cultural similarities and differences in parenting. *Journal of Child Psychology and Psychiatry*, 63, 466–479.
- Leahy, B., & Carey, S. E. (2020). The acquisition of modal concepts. *Trends in Cognitive Sciences*, 24, 65–78.

- Leahy, B., & Zalnieriunas, E. (2021). Might and might not: Children's conceptual development and the acquisition of modal verbs. In *Semantics and linguistic theory* (pp. 426–445).
- Liszkowski, U., Carpenter, M., & Tomasello, M. (2008, September). Twelve-month-olds communicate helpfully and appropriately for knowledgeable and ignorant partners. *Cognition*, 108, 732–739.
- Liu, S., & Spelke, E. S. (2017). Six-month-old infants expect agents to minimize the cost of their actions. *Cognition*, 160, 35–42.
- Liu, S., Ullman, T. D., Tenenbaum, J. B., & Spelke, E. S. (2017). Ten-month-old infants infer the value of goals from the costs of actions. *Science*, 358, 1038–1041.
- Magid, R. W., DePascale, M., & Schulz, L. E. (2018). Four-and 5-year-olds infer differences in relative ability and appropriately allocate roles to achieve cooperative, competitive, and prosocial goals. *Open Mind*, 2, 72–85.
- Martin, A., & Olson, K. R. (2013). When kids know better: paternalistic helping in 3-year-old children. *Developmental Psychology*, 49, 2071.
- McElreath, R. (2018). *Statistical rethinking: A bayesian course with examples in r and stan*. Chapman and Hall/CRC.
- McGhee, P. E. (1974). Moral development and children's appreciation of humor. *Developmental Psychology*, 10, 514.
- McGhee, P. E., Goldstein, J. H., et al. (1983). *Handbook of humor research* (Vol. 1). Springer.
- Meyer, M., van der Wel, R. P., & Hunnius, S. (2016). Planning my actions to accommodate yours: joint action development during early childhood. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371, 20150371.
- Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic bulletin & review*, 23, 103–123.
- Nicenboim, B., Schad, D., & Vasishth, S. (2021). An introduction to bayesian data analysis for cognitive science. *Under contract with Chapman and Hall/CRC statistics in the social and behavioral sciences series*.
- Noveck, I. A. (2001). When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition*, 78, 165–188.
- Opie, I. A., & Opie, P. (2001). *The lore and language of schoolchildren*. New York Review of Books.
- Pinker, S., Nowak, M. A., & Lee, J. J. (2008). The logic of indirect speech. *Proceedings of the National Academy of sciences*, 105, 833–838.

- Powell, L. J. (2022). Adopted utility calculus: Origins of a concept of social affiliation. *Perspectives on Psychological Science*, 17, 1215–1233.
- Qian, P., Bridgers, S., Taliaferro, M., Parece, K., & Ullman, T. D. (2024). Ambivalence by design: A computational account of loopholes. *Cognition*, 252, 105914.
- Rafetseder, E., Cristi-Vargas, R., & Perner, J. (2010). Counterfactual reasoning: Developing a sense of “nearest possible world”. *Child development*, 81, 376–389.
- Rafetseder, E., Schwitalla, M., & Perner, J. (2013). Counterfactual reasoning: From childhood to adulthood. *Journal of experimental child psychology*, 114, 389–404.
- Schlenker, B. R., & Weigold, M. F. (1992). Interpersonal processes involving impression regulation and management. *Annual review of psychology*, 43, 133–168.
- Shultz, T. R., & Horibe, F. (1974). Development of the appreciation of verbal jokes. *Developmental Psychology*, 10, 13.
- Silveira, F., Shafer, K., Dufur, M. J., & Roberson, M. (2021). Ethnicity and parental discipline practices: A cross-national comparison. *Journal of Marriage and Family*, 83, 644–666.
- Skordos, D., & Papafragou, A. (2016). Children’s derivation of scalar implicatures: Alternatives and relevance. *Cognition*, 153, 6–18.
- Snyder, C. R., & Higgins, R. L. (1988). Excuses: Their effective role in the negotiation of reality. *Psychological bulletin*, 104, 23.
- Sommerville, J. A., Enright, E. A., Horton, R. O., Lucca, K., Sitch, M. J., & Kirchner-Adelhart, S. (2018). Infants’ prosocial behavior is governed by cost-benefit analyses. *Cognition*, 177, 12–20.
- Stiller, A. J., Goodman, N. D., & Frank, M. C. (2015). Ad-hoc implicature in preschool children. *Language Learning and Development*, 11, 176–190.
- Struchiner, N., Hannikainen, I. R., & de Almeida, G. d. F. (2020). An experimental guide to vehicles in the park. *Judgment and Decision Making*, 15, 312–329.
- Svetlova, M., Nichols, S. R., & Brownell, C. A. (2010). Toddlers’ prosocial behavior: From instrumental to empathic to altruistic helping. *Child development*, 81, 1814–1827.
- Tomasello, M. (2009). *Why we cooperate*. MIT Press, Cambridge, MA.
- Tomasello, M. (2018). How children come to understand false beliefs: A shared intentionality account. *Proceedings of the National Academy of Sciences*, 115, 8491–8498.
- Ullman-Margalit, E. (1983). On presumption. *The Journal of Philosophy*, 80, 143–163.
- Uther, H.-J. (2004). *The types of international folktales—a classification and bibliography*. Suoma-

- lainen Tiedekatemia Academia Scientiarum Fennica Exchange Centre.
- Warneken, F., Steinwender, J., Hamann, K., & Tomasello, M. (2014). Young children's planning in a collaborative problem-solving task. *Cognitive Development*, 31, 48–58.
- Warneken, F., & Tomasello, M. (2006). Altruistic helping in human infants and young chimpanzees. *Science*, 311, 1301.
- Warneken, F., & Tomasello, M. (2007). Helping and cooperation at 14 months of age. *Infancy*, 11, 271-294.
- Winner, E., Levy, J., Kaplan, J., & Rosenblatt, E. (1988). Children's understanding of nonliteral language. *Journal of Aesthetic Education*, 22, 51–63.
- Woo, B. M., Liu, S., Gweon, H., & Spelke, E. S. (2024). Toddlers prefer agents who help those facing harder tasks. *Open Mind*, 8, 483–499.
- Woo, B. M., & Spelke, E. S. (2023). Infants and toddlers leverage their understanding of action goals to evaluate agents who help others. *Child Development*, 94, 734–751.