Title

Loopholes, a Window into Value Alignment and the Learning of Meaning

Permalink

https://escholarship.org/uc/item/3q48s73j

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 43(43)

ISSN

1069-7977

Authors

Bridgers, Sophie
Schulz, Laura
Ullman, Tomer D.

Publication Date

2021

Peer reviewed

Loopholes, a Window into Value Alignment and the Learning of Meaning

Sophie Bridgers (secb@mit.edu)
Brain and Cognitive Sciences Department, Massachusetts Institute of Technology
Department of Psychology, Harvard University

Laura E. Schulz (lschulz@mit.edu)
Brain and Cognitive Sciences Department, Massachusetts Institute of Technology

Tomer D. Ullman (tullman@fas.harvard.edu)
Department of Psychology, Harvard University
Cambridge, MA, USA

Abstract

Finding and exploiting loopholes is a familiar facet of fable, law, and everyday life. But cognitive, computational, and empirical work on this behavior remains scarce. Engaging with loopholes requires a nuanced understanding of goals, social ambiguity, and value alignment. We trace loophole behavior to early childhood, and we propose that exploiting loopholes results from a conflict in actors' goals combined with a pressure to cooperate. A survey of 260 parents reporting on 425 children reveals that loophole behavior is prevalent, frequent, and diverse in daily parent-child interactions, emerging around ages five to six and tapering off from around ages nine to ten into adolescence. A further experiment shows that adults consider loophole behavior in children as less costly than non-compliance, and children increasingly differentiate loophole behavior from non-compliance from ages four to ten. We discuss limitations of the current work together with a proposal for a formal framework for loophole behavior.

Keywords: pragmatic communication; utility trade-offs; cooperation; development; loopholes

A child plays with toys scattered in her room. Her mother enters and says, "When I come back, I don't want to see *anything* on the floor." Not wanting to put her toys away, but worried about the consequences of disobedience, the child finds herself in a dilemma. With a stroke of insight, she finds a solution. When the mother comes back, she finds that the toys are all in a heap on top of the bed. The toys are still available for play, and her instructions were met. Technically.

While potentially low-stakes and humorous to the adult eye, this everyday example makes clear two central challenges of human cooperation that have wide-ranging implications: goal communication, and goal alignment. Conveying goals and inferring the goals of others are complex processes, as utterances are ambiguous, and a single behavior may be consistent with any of a rich space of possible goals. Even if we do reasonably recover what someone else wants from us, we still face the decision of whether to comply. Our goals often don't align perfectly with others', but refusing to help or cooperate can be costly—we could irritate or upset our social partner; they might even retaliate or exact punishment.

In cases of misalignment, the ambiguity of language can provide an opening, a *loophole*. Between compliance and refusal there exists a vast gray area where people can feign confusion, intentionally misunderstand, obey the letter of the law but not the spirit, do what was asked but not what was wanted, and so on. Acting on a loophole can save an agent from giving up on their own goals but also can reduce the likely social retribution of outright non-compliance with another's goal. People can resolve the second challenge of cooperation by exploiting the first.

Loophole-seeking is a familiar phenomenon. There is an entire area of law devoted to "malicious compliance", and perennial concern with 'form vs. substance', and 'letter vs. spirit of the law' distinctions (Isenbergh, 1982; Katz, 2010). In history, intentional misunderstandings have been used by workers, soldiers, and other populations who could not stand to obey, but could not risk to disobey (Scott, 1985). In art and fable, centuries-old stories of outwitting malevolent forces through clever misinterpretations, or being similarly tricked by a mischievous spirit, appear often enough to form separate sub-genres (Uther, 2004). And of course, willful misunderstanding is a hallmark of childhood (e.g., in games of guile; Opie & Opie, 2001). Yet, to our knowledge there is no detailed cognitive, computational, or developmental study of how humans learn to find these creative workarounds.

Previous research has largely focused on how humans learn to communicate and act cooperatively (Bohn & Frank, 2019; Tomasello, 2009). Loopholes, however, subvert the usual process of goal inference and joint action. Loopholes offer a different lens for the typical workings of cooperation and reasoning about intention, in the same way that visual illusions help to shed light on the implicit assumptions and computations made by the visual system. Understanding the emergence of loophole behavior in childhood, specifically, can uncover the representations that support it, as children may learn to find and exploit loopholes as a natural part of their developing understanding of communication and cooperation. The drive and ability to help and understand others emerges early (e.g., Gergely & Csibra, 2003; Warneken & Tomasello, 2006), but a deeper comprehension of goals and ambiguity that enables one to leverage the under-specification of social interaction for one's own gain may emerge later in childhood.

We suggest that the ability to reason about loopholes requires a complex integration of an understanding of pragmatics, utilities, and joint-planning. Pragmatics is understanding language in context. It involves reasoning about the speaker's intentions to disambiguate the intended meaning from the space of plausible alternatives (e.g., Bates, 1976; Goodman & Frank, 2016). Although children use an understanding of others' goals to learn language starting in infancy, the comprehension of meaning beyond literal content undergoes sub-

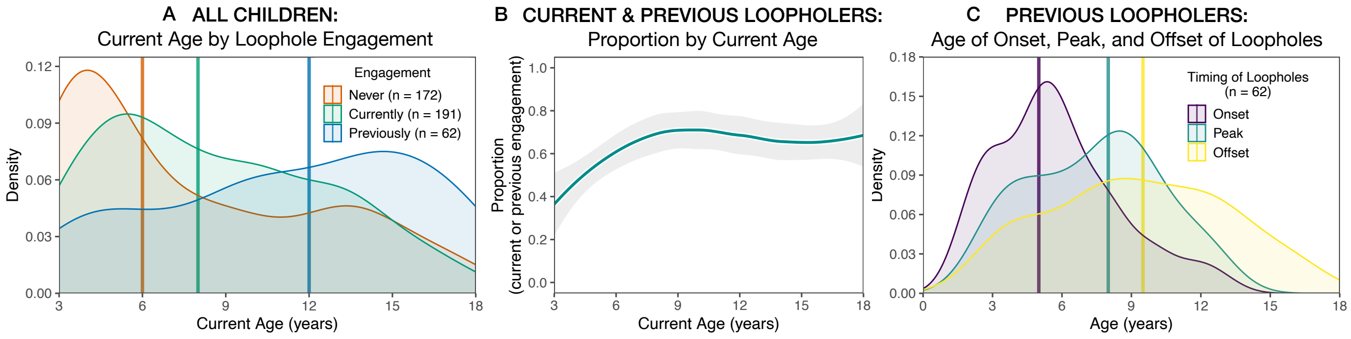

Figure 1: **Study 1. A** Distribution and median age of children reported to have engaged never (red), currently (green), or previously (blue) in loopholes. **B** Current age by proportion of current/previous loophole engagers ("loopholers") of that age (95% CI in grey). **C** Distribution and median age of loophole onset (purple), peak (turquoise), and offset (yellow) for previous loophole engagers.

stantial development throughout childhood (Bohn & Frank, 2019). In particular, young children may struggle with identifying relevant alternatives to the literal utterance (e.g., Barner, Brooks, & Bale, 2011). Irony, metaphor, puns, and sarcasm further complicate the process of honing in on intended meaning (Winner, Levy, Kaplan, & Rosenblatt, 1988). Although children as young as five are able to reject the literal meaning of many utterances, the ability to understand communicative intent continues to develop into adolescence (Demorest, Silberstein, Gardner, & Winner, 1983).

Much like indirect or non-literal language, finding a loophole involves representing multiple possible interpretations of an utterance. The listener can then deliberately act on an unintended but possible interpretation instead of following the speaker's intended meaning. We might expect these abilities to emerge around five to seven years of age, correlating with the development of related abilities including higher order Theory-of-Mind (e.g., knowing what someone believes about someone else's beliefs) and explicit comparison of the probabilities of different events (Filippova & Astington, 2008; Leahy & Carey, 2020; Tomasello, 2018).

But understanding loopholes requires more than representing an alternative meaning for an utterance; the listener also has to understand that an intended meaning may conflict with their own goals and that another possible but prima facie less plausible meaning may better serve their own ends. Such reasoning requires representing the utilities of the speaker and listener, and the costs of both complying with and refusing the speaker's request.

Children begin to reason about the utilities (costs and rewards) of others' actions early in development (Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016). Between four and seven, they can use estimates of others' expected utilities in sophisticated ways, such as resolving pragmatic ambiguity (Jara-Ettinger, Floyd, Huey, Tenenbaum, & Schulz, 2020). And although children show an early willingness to help others, this motivation is mitigated by the physical or resource costs of responding (e.g., Sommerville et al., 2018; Svetlova, Nichols, & Brownell, 2010).

In this prior research, the influence of personal costs was only measured by whether or not children responded to a request for help, and there was implicit social pressure but no real consequence for not helping. Thus we know little empirically about how children handle situations where they do not want to comply with a request but cannot easily refuse. In third party judgments, however, children become increasingly likely to accept violations of a moral rule when doing so has positive consequences (e.g., Neary & Friedman, 2014). Conversely (but drawing on similar cognitive capacities), children between four and ten become increasingly likely to think that someone who violated the letter of a rule but upheld the spirit should be treated more leniently than those who violated both the letter and spirit (Bregant, Wellbery, & Shaw, 2019).

Collectively, the research on children's understanding of ambiguous language, conflicting utilities, and rule violations suggests that we should expect to see an understanding of loopholes emerge around age five, and continue to develop through middle childhood. In Study 1, we survey parents to gather reports about the emergence and prevalence of loophole behavior in naturalistic settings. In Study 2, we test the hypothesis that in certain situations, loophole behavior does indeed serve the role of allowing the listener to achieve her own ends while mitigating the costs of refusing a speaker who has an opposing goal. We test both adults (Study 2a) and children (Study 2b) to see if participants predict that a listener who engages in loophole behavior will get in less trouble than one who is simply non-compliant. We end by proposing a novel computational framework of goal communication that supports loophole behavior, and by discussing the implications of this research for improved insight into both human communication, and safer human-AI interactions.

Study 1: How pervasive are loopholes, and when do they emerge?

We surveyed parents to study the emergence, extent, and scope of loophole behavior in childhood.

Participants. Participants were 260 parents of children between the ages 3 and 18 years (inclusive), recruited online via Prolific. The survey took approximately 9 minutes and 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; 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. Participants read a definition of loophole behavior, including examples of children finding loopholes in parents' requests. Participants then classified loophole vs. non-compliant behaviors in two stories. They were then asked to report for each of their own children: (1) current age, (2) whether they currently engage, used to engage, or never engaged with loopholes, and where applicable: (3) onset, peak frequency, and offset age of loophole behavior, and (4) how frequently this behavior occurs. Parents were also invited to share examples of their children's loophole behavior.

Results. Parents readily understood what was meant by loophole behavior (93% correctly identified it; 91% correctly identified non-compliance), and many parents recalled specific instances of such behavior in their own children. A majority of children (60%) were reported as engaging in loophole behavior currently (45%) or previously (15%) (Fig. 1A). According to parent report, children begin engaging with loopholes at 5 to 6 years (M_{age}: 5.6, range: 2 to 13 yrs), do so most frequently at ages 7-8 (M_{age}: 7.4, range: 2 to 13 yrs), and taper off around ages 9-10 (M_{age}: 9.3, range: 3 to 17 yrs) (Fig. 1B-C). When children are engaging with loopholes, they do so regularly (once every few days to few weeks).

Our survey indicates that loophole behavior: (1) is easily recognized by parents; (2) is prevalent and frequent in parent-child interactions; (3) emerges around an age (5-6 years) that corresponds to increased sophistication in pragmatics and Theory-of-Mind (e.g., Barner et al., 2011; Tomasello, 2018); (4) is a general cognitive phenomenon and not specific to particular linguistic constructions or conceptual domains: Parents shared rich anecdotes of how children found loopholes with scalars, timing, scope, reference, knowledge, and more (e.g., parent: "You need to do some reading," child reads a sentence; parent: "Stop jumping on that couch," child switches to jumping on the other couch).

Study 2: How do children and adults evaluate loophole behavior?

Study 1 established loopholes as an ecologically valid behavior in childhood based on parent report. The survey, however, reflected the parents' point-of-view. What do children understand about loopholes? We hypothesized that one possible function of loopholes is that they maximize utilities by allowing people to achieve their own goals while reducing social costs. For example, a child who exploits a loophole in their parent's request might get into less trouble than if they had outright not complied because the request was technically met, and there is some plausible deniability (i.e., the child could feign genuine confusion). What's more, the behavior is clever and might even be funny to the parent since humor can be a function of violated expectations. Indeed, some parents shared comments stating that loophole behavior made them laugh, attributing it to "added brainpower," and explaining that they couldn't really punish their child because "...people do it all the time and...there are times where they should take advantage of loophole behavior."

We empirically tested the proposal that loopholes can be less socially costly than non-compliance by examining whether adults and children estimate that loopholes decrease the likely degree of punishment and parental upset compared to non-compliance, as well as increase likely amusement.

Study 2a: Adults' reasoning about the consequences of children's loophole behavior

Participants. Participants (N = 55; M_{age}: 32.5, range: 18 to 65 yrs, 55% female, 2% non-binary) with a 95% approval rating, who lived in the U.S., and were fluent in English were recruited online via Prolific. The survey took approximately 8 minutes, and compensation was $1.43. Participants were majority White (64%; 11% Black, 11% Hispanic or Latinx, 7% Asian, 4% multi-racial) from diverse regional and educational backgrounds. An additional 5 participants were recruited but excluded from analysis due to failing an attention check.

Procedure. We created 27 different scenarios (9 stories with 3 endings each) based on real-life examples provided in Study 1. Each scenario had a parent who made a request of a child (e.g., "Put down the tablet"), and the child either complied, did not comply, or found a loophole (e.g., child puts tablet down but keeps looking at it). Participants read nine scenarios (3 compliance, 3 non-compliance, 3 loophole) in a Qualtrics survey. The order of the scenarios and the condition (ending) of each scenario were counterbalanced across participants. Participants were informed that they would (1) read nine scenarios each about a child who is asked to do something by their parent and then responds in some way, and (2) be asked about the consequences of the child's response.

For each scenario, participants evaluated the child's response on a 4-point scale according to (1) how much trouble the child would get into (*no trouble / a little bit of trouble / trouble / a lot of trouble*), (2) how upset the parent would be (*not upset / ... / very upset*), and (3) how funny the parent would find the behavior (*not funny /.../ very funny*). Participants responded by filling in the blank of three sentences (order counterbalanced across participants) with a phrase from a drop-down menu (e.g., selecting level of upset for "Avni's mother feels_____about what Avni is doing."). The survey can be viewed here: `https://harvard.az1.qualtrics.com/jfe/form/SV_7TB82SEtLi7xEsS`

Results. We conducted a mixed effects linear regression predicting adults' ratings of the degree of trouble, upset, and funniness on a 4-point scale (coded as an integer from 0-3) with main effects of condition (3-levels: compliance, loophole, non-compliance) and measure (3-levels: trouble, upset, funny), as well as their interaction with the maximal ran-

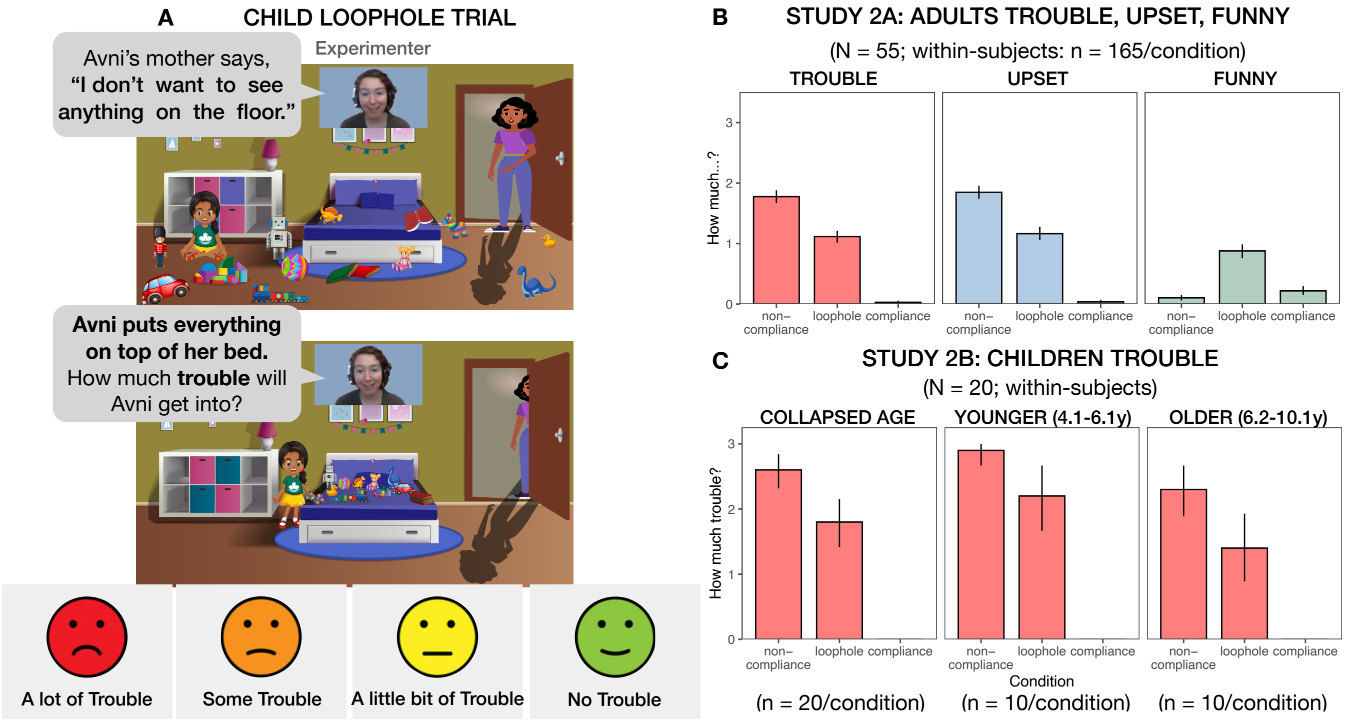

Figure 2: **Study 2 Procedure and Results.** **A** Example loophole scenario and trouble scale for children. **B** Adults' ratings of trouble (red), upset (blue), and funniness (green) on 4-point scale for children's non-compliance (left-bar), loophole-seeking (middle-bar), and compliance (right-bar). **C** Children's ratings of trouble: Collapsed Age, Younger, and Older (median age split). Error bars: 95% bootstrapped CIs.

dom effects structure that converged (random intercepts and effects of condition and measure by subject and scenario). Adults distinguished loophole behavior from compliance and non-compliance: they believed it would result in the child getting into less trouble and the parent being less upset than non-compliance (trouble: $\beta = -0.65, SE = 0.09, t = -7.24$, upset: $\beta = 0.68, SE = 0.09, t = 7.50$) and would be more amusing than compliance ($\beta = 0.65, SE = 0.10, t = 6.46$) or non-compliance ($\beta = 0.77, SE = 0.09, t = 8.54$). (Fig. 2B)

Study 2b (pilot): Children's reasoning about loophole behavior

Participants. Due to recruitment constraints during the pandemic, we used a convenience sample of children in the U.S. and the U.K. (N = 20; M_{age}: 6.7, range: 4.1 to 10.1 yrs, 40% female, all White) tested online over Zoom, and thus we consider this study a pilot experiment.

Procedure. We selected five of the nine scenarios used with adults in Study 2a and had them illustrated by an artist (see `https://osf.io/rwgmx/` for details). Children saw three of these scenarios (one loophole, one compliance, and one non-compliance) presented as novel story-books displayed over Zoom and narrated aloud by an experimenter. Which scenarios children saw, the condition of each scenario, and the order of the conditions were counterbalanced across participants.

Children 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. For each scenario, children indicated how much trouble the child protagonist would get into on a 4-point scale, with each point represented as a different colored face expressing a different affect (Fig 2A). Children received training and practiced using the scale ahead of time. Children were also asked to explain their choice of trouble. As exploratory measures, we coded children's own amusement upon hearing the child protagonist's response (indexed by whether they smiled or laughed), and asked them to compare the non-compliant and loophole protagonists' parents in terms of who was more upset and who was more amused.

Results. We conducted a mixed effects linear regression predicting children's ratings of the degree of trouble on a 4-point scale (integer from 0-3) with main effects of condition (3-levels: compliance, loophole, non-compliance) and age-group (2-levels: younger, older determined by a median age split), as well as their interaction with random intercepts by subject and random intercepts and effects of condition and age-group by scenario. Similar to adults, children thought loophole behavior would result in less trouble than non-compliance (4.1-6.1 years: $\beta = 0.61, SE = 0.22, t(8.73) = 2.76, p = .023$; 6.2-10.1 years: $\beta = 1.16, SE = 0.26, t(11.76) = 4.46, p < .001$), with suggestive evidence that this distinction was greater for older than younger children ($\beta = 0.55, SE = 0.32, t(20.97) = 1.74, p = .096$). Older children also rated loophole behavior as resulting in less trouble than younger children ($\beta = 0.96, SE = 0.28, t(14.24) = 3.49, p = .004$). (Fig. 2C)

Exploratory observations suggested that children's explanations of trouble differed for non-compliance vs. loopholes: for non-compliance, children appealed to not listening (e.g., "He didn't do what his mother said"). For loopholes, older children identified ambiguity in the request (e.g., "It is doing what her mom said but not exactly what her mom meant"), and described the child as trying to trick or "get around" the parent. Older children thought the non-compliant child's parent would be more upset (of children who responded, 7/10 of older vs. 3/8 of younger). Both older and younger children thought parents would find loopholes funnier than non-compliance (of children who responded, 8/8 of older, 4/6 of younger) and smiled or laughed more themselves for loopholes (8/20 loophole, 1/20 compliance, 1/20 non-compliance). These observations are based on small numbers, but we speculate that in addition to trouble, children may distinguish loopholes from (non)compliance in terms of humor and parental upset, and may be explicitly aware of the ambiguity that loopholes exploit.

General Discussion

Previous cognitive research has largely focused on how humans learn to communicate and act cooperatively (e.g., Bohn & Frank, 2019; Tomasello, 2009), but not on how people skirt cooperativeness and willfully pursue possible but unintended interpretations of others' goals. We presented two studies that, to our knowledge, are the first to systematically explore (1) the emergence of loophole behavior in parent-child interactions and (2) children's and adults' intuitions about the function of loopholes in these interactions. We find that loophole behavior is prevalent and diverse in childhood, emerging around ages 5-6, and peaking around 7-8 before tapering off into adolescence (Study 1). We hypothesized that loopholes are a means to achieve one's own goals while reducing the probability or severity of social penalty. Both adults' and children's evaluations of loopholes vs. non-compliance were consistent with this hypothesis (Study 2). From ages four to ten, children were increasingly likely to believe that exploiting a loophole would result in less trouble, paralleling the developmental trajectory of loophole behavior in Study 1 and suggesting children's ability to distinguish others' loophole vs. non-compliant behavior may correlate with the degree to which they exploit loopholes themselves. Some of the eldest children in Study 2 even spontaneously identified the tension between conflicting goals and social pressure underlying loophole-seeking (e.g., 10yo: "He kind of found a way to still do what he wanted to do but around his dad's command"; 9yo: "He could lie about being confused about what his dad meant and get to play more video games").

This work is a first step in a more detailed empirical and formal study of the development of loophole behavior. In the rest of the discussion, we consider limitations and open questions in the current data, together with planned extensions for future work. We then sketch a proposal for a formal framework for loophole behavior grounded in Rational Speech Acts (RSA) models, a leading framework for cooperative pragmatic communication (Goodman & Frank, 2016). Lastly, we consider this work in the context of the pressing issue of value alignment in human-machine interactions.

The parent survey from Study 1 presented a possible developmental trajectory for loophole behavior. At the same time, there is quite a bit of variance in parent reported age of onset, peak, and offset. Parent report is informative but also limited as it relies both on parents' memory and ability to correctly identify loophole behavior as distinct from genuine confusion and from non-compliance. Indeed, some parents reported that their children started to exploit loopholes at age two, which seems unlikely and more probable that the children honestly misunderstood what their parents wanted. Children's responses in Study 2b are consistent with the idea that loophole understanding emerges between four and ten years of age, but this is a preliminary and exploratory study.

In order to more robustly and precisely interrogate the developmental trajectory of the understanding of loopholes, we first need to scale up and extend our pilot Study 2b. In ongoing work, we are replicating this study with a larger, more diverse participant sample and more scenarios to test the hypothesis that children by ages 5-6 (but not earlier) differentiate loopholes from non-compliance and compliance. We are also exploring other cognitive capacities (e.g., Theory-of-Mind, pragmatic reasoning, executive function) that might correlate with the emergence of loophole understanding, as well as children's own loophole engagement. This work will lend insight to the connection between loophole comprehension and production, as well as the representations and inferential machinery that support this behavior.

We proposed that exploiting loopholes minimizes the repercussions that would result from failing to comply with someone's request because there would exist a possible interpretation under which the request was fulfilled (and the behavior might be considered clever or funny at least among parents and children). But people don't evaluate others' behavior based solely on its degree of compliance; they also consider the outcome and the underlying intentions (e.g., Cushman, 2008; Cushman, Sheketoff, Wharton, & Carey, 2013). In Study 2, we did not not provide nor query participants about the protagonist's intentions. Nonetheless, some children appealed to the protagonist's latent mental states to justify the amount of trouble they would get into. As one 9-year-old explained (when the protagonist found a loophole in the request to stop playing on Xbox by picking up a PlayStation), the protagonist would get into "no trouble because he might have been confused and thought it was just the Xbox or a little bit of trouble because maybe he knows and is deliberately trying to do it." Other children also indicated that legitimate confusion should be treated more leniently than feigning confusion. It is possible that some children may even have interpreted the loophole protagonist as genuinely (rather than intentionally) misunderstanding the parent. Participants' judgments also appeared sensitive to outcome: for the same scenario, loopholes were rated as less problematic than non-compliance, but across scenarios, ratings for loopholes and non-compliance varied. For example, in one scenario the child is told not to go outside alone, and adults thought both the non-compliant child (who went outside alone) and the loophole child (who went outside with the dog) would get into more trouble than the non-compliant child and loophole child told not to eat all of the popcorn, presumably because going outside without an adult could lead to worse outcomes than eating too much.

Loopholes, however, may not always be less costly than non-compliance even in the same situation. For example, a child who is told to stop hitting their sister and so then starts to kick their sister might get into even more trouble than a child who continues hitting their sister. Also, if both parties know that the listener understood what was intended, then the plausible deniability is no longer on the table and so the behavior might be treated less generously. This might be one reason why loophole behavior begins to taper off: as children get older, genuine misunderstanding is less plausible and the behavior has lost its novelty and so might be viewed as less funny or clever. Of course, an individual might exploit a loophole precisely in order to annoy or upset someone (i.e., the loophole might be genuinely malicious rather than playful or mischievous).

Future work will look at how adults and children integrate information about intentions and outcomes with evaluations of compliant, non-compliant, and loophole behavior. We will also interrogate adults' and children's intuitions about the motivations that underlie loophole behavior and look at when they attribute actual loophole-seeking vs. genuine misunderstanding, as well as malignant vs. more benign intent. Finally, we plan to empirically manipulate the costs of compliance and non-compliance, as well as the ambiguity of language to see if adults and children respond systematically to these variations in predicting, explaining, and evaluating loophole-seeking. This work will provide the basis for our formal framework described below.

Proposal for formal framework of loopholes. We aim to extend the cooperative RSA framework to better understand how individuals' conflicting goals can give rise to intentional misunderstanding. A loophole-seeking individual needs to understand (1) what is being asked (i.e., the speaker's goals), (2) what their own goals are (i.e., the listener's goals), and (3) what the trade-off is between their goals and the speaker's goals. We propose to formalize these ingredients by synthesizing cooperative RSA models (what is being asked), with rational planning models (what are my goals), and utility trade-off frameworks (value alignment).

In a standard **RSA** set-up, a speaker and listener collaborate to reason about a space of intended meanings (Goodman & Frank, 2016). Given a specific utterance, the listener considers a speaker who is reasoning about a listener and computes the distribution over intended meaning. The speaker's utility is typically linked to whether the listener correctly infers the intended meaning. Our framework will combine this RSA framework with **planning frameworks**, specifically expected utility maximization (e.g. Russell & Norvig, 2020; Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017). In our framework, the intended meaning is itself the speaker's utility (goal). The listener chooses actions to maximize their own utility, but the listener's utility can take into account the speaker's utility, leading to collaborative or helpful acts through **joint planning** (cf. Ullman et al., 2009; Russell & Norvig, 2020). We suggest that standard shallow goal-communication pipes the speaker's utility (the output of RSA) into planning to produce an action. However, in deeper goal understanding, after a first pass of decoding the speaker's utility by RSA, a low utility outcome for the listener may be predicted by planning. This can trigger a 'loophole search', in which the product of possible interpretations of RSA are re-weighted by their usefulness. A useful unintended meaning can be 'supposed' and fed into planning (cf. suppositions in imagination Harris, 2000). Key loci for developmental change include trigger functions, generation of meaning spaces, supposition, and joint planning. We will compare human behavior in future experiments to different versions of the model that have or lack key components, creating a formal framework that generates hypotheses for how goal communication grows into adult understanding.

Value misalignment in human-technology interactions. The complexity of goal communication is not only reflected in human loophole behavior, but also in engineered systems that 'do what you say, but not what you want.' People struggle to explicitly specify their full intended values and desires, leading to machines that achieve high performance on a measure that has nothing to do with the task (e.g., algorithms learning to deliberately delete games in order to avoid the negative score of losing; Krakovna, 2020). This misbehavior is not due to a particular sort of algorithm, and many documented failures exist across methods and domains (Lehman et al., 2020). Taken to the extreme, the problems of value alignment pose significant risks for human-machine interactions and have become a major concern among researchers and policy makers (Amodei et al., 2016). Current machines do not willfully misunderstand goals any more than a bridge is being lazy by falling down. But a better understanding of the psychological processes that let even young humans intuitively solve and purposefully contort goal communication could inform the design of safer intelligent machines.

Loopholes are pervasive, consequential, and offer a unique window into the commonsense process of goal understanding and how people navigate the tension between their own and others' goals—central challenges to successful cooperation. The current work offers a foundation for developmentally and computationally characterizing loopholes, supporting new frameworks for analyzing social decision making.

Acknowledgments

We thank the families who participated in this research, and the members of the MIT Early Childhood Cognition Lab, members of the Harvard Computation, Cognition, and Development Lab, as well as Drs. Julia Leonard, MH Tessler, and Natalia Velez for their helpful comments and discussion. This research is funded by a MIT Simons Center for the Social Brain Postdoctoral Fellowship (SB) and the MIT Center for Brains, Minds, and Machines (TU).

References

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*.

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*, 0064.

Barner, D., Brooks, N., & Bale, A. (2011). Accessing the unsaid: The role of scalar alternatives in children's pragmatic inference. *Cognition*, *118*(1), 84–93.

Bates, E. (1976). *Language and context: The acquisition of pragmatics*. Academic Press.

Bohn, M., & Frank, M. C. (2019). The pervasive role of pragmatics in early language. *Annual Review of Developmental Psychology*, *1*, 223–249.

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.

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

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

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*(2), 121–134.

Filippova, E., & Astington, J. W. (2008). Further development in social reasoning revealed in discourse irony understanding. *Child Development*, *79*(1), 126–138.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: the naïve theory of rational action. *Trends in Cognitive Sciences*, *7*(7), 287–292.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11).

Harris, P. L. (2000). *The work of the imagination.* Blackwell Publishing.

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*(8), 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*(4), 1135–1149.

Katz, L. (2010). A theory of loopholes. *The Journal of Legal Studies*, *39*(1), 1–31.

Krakovna, V. (2020). *Specification gaming examples in AI - master list.* `http://bit.ly/kravokna_examples_list.` (Accessed: 2020-12-28)

Leahy, B. P., & Carey, S. E. (2020). The acquisition of modal concepts. *Trends in Cognitive Sciences*, *24*(1), 65–78.

Lehman, J., Clune, J., Misevic, D., Adami, C., Altenberg, L., Beaulieu, J., … others (2020). The surprising creativity of digital evolution: A collection of anecdotes from the evolutionary computation and artificial life research communities. *Artificial Life*, *26*(2), 274–306.

Neary, K. R., & Friedman, O. (2014). Young children give priority to ownership when judging who should use an object. *Child Development*, *85*(1), 326–337.

Opie, I. A., & Opie, P. (2001). *The lore and language of schoolchildren*. New York Review of Books.

Russell, S., & Norvig, P. (2020). *Artificial intelligence: A modern approach* (4th ed.). Pearson.

Scott, J. C. (1985). *Weapons of the weak: Everyday forms of peasant resistance*. yale university Press.

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.

Svetlova, M., Nichols, S. R., & Brownell, C. A. (2010). Toddlers' prosocial behavior: From instrumental to empathic to altruistic helping. *Child Development*, *81*(6), 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*(34), 8491–8498.

Ullman, T. D., Baker, C. L., Macindoe, O., Evans, O., Goodman, N. D., & Tenenbaum, J. B. (2009, October). Help or hinder: Bayesian models of social goal inference. *Advances in Neural Information Processing Systems*, 1874–1882.

Uther, H.-J. (2004). *The types of international folktales–a classification and bibliography*. Suomalainen Tiedeakatemia Academia Scientiarum Fennica Exchange Centre.

Warneken, F., & Tomasello, M. (2006). Altruistic helping in human infants and young chimpanzees. *Science*, *311*(5765), 1301.

Winner, E., Levy, J., Kaplan, J., & Rosenblatt, E. (1988). Children's understanding of nonliteral language. *Journal of Aesthetic Education*, *22*(1), 51–63.