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

Sophie Bridgers
Department of Brain & Cognitive Sciences
Massachusetts Institute of Technology
Cambridge, MA 02139
secb@mit.edu

Elena Glassman
School of Engineering and Applied Sciences
Harvard University
Cambridge, MA 02138
glassman@seas.harvard.edu

Laura Schulz
Department of Brain & Cognitive Sciences
Massachusetts Institute of Technology
Cambridge, MA 02139
lschulz@mit.edu

Tomer Ullman
Department of Psychology
Harvard University
Cambridge, MA 02138
tullman@fas.harvard.edu

Abstract

Exploiting a loophole, taking advantage of the ambiguity of language to do what someone says but not what they want, is a familiar facet of fable, law, and everyday life. Engaging with loopholes requires a nuanced understanding of goals, social ambiguity, and value alignment. Scientifically, the development of loopholes can help us better understand human communication, and design better human-AI interactions. However, cognitive research on this behavior remains scarce. A survey of parents reveals that loophole behavior is prevalent, frequent, and diverse in daily parent-child interactions, emerging around ages five to six. A further experiment shows that adults consider loophole behavior as less costly than non-compliance, and children increasingly differentiate loophole behavior from non-compliance from ages four to ten. We discuss the implications and limitations of the current work, together with a proposal for a formal framework for loophole behavior.

1 Introduction

A father tells his daughter, "It's time to put the tablet down." Not wanting to stop using the tablet, but worried about the consequences of disobedience, the child finds herself in a dilemma. With a stroke of insight, she puts the tablet down on the table in front of her, and keeps playing with it. She can still use the tablet, and her father's instructions were met. Technically.

While potentially low-stakes and humorous to the adult eye, this everyday example highlights two central challenges of cooperation: 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 many possible goals. And even if we 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. But in these 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, obey the letter of the law but not the spirit, do what was asked but not what was wanted, and so on.

Meaning in Context: Pragmatic Communication in Humans and Machines Workshop at 35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

Loophole-seeking is a familiar facet of human society and everyday life. Willful misunderstanding is a hallmark of childhood (e.g., in games of guile). (1) In law, there is perennial concern with "malicious compliance", as well as with 'form vs. substance' and 'letter vs. spirit of the law' distinctions. (2; 3) Across history, intentional misunderstandings have been used by populations who could not stand to obey, but could not risk to disobey. (4) And in art and fable, there are centuries-old stories of people outwitting malevolent forces through clever misinterpretations, or being similarly tricked by mischievous spirits. (5)

While malevolent spirits aren't a current issue, the possibility of rogue AI's misinterpreting goals and causing unintended harm while obeying their technical specifications has become a pressing concern among researchers and policy makers. (6; 7) Engineers 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). (8) This misbehavior is not due to a particular sort of algorithm, and many documented failures exist across methods and domains. (9) 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 in the future.

Understanding the emergence of loophole behavior in childhood can uncover the representations that support it, as exploiting loopholes may be a natural part of children's developing understanding of communication and cooperation. The drive and ability to understand and help others emerges early (10; 11; 12; 13), but a deeper comprehension of goals, ambiguity, and utility trade-offs that enables one to leverage the under-specification of social interaction for one's own gain may emerge later in childhood. From ages five to seven, children explicitly reason about the costs and rewards of others' actions (14) and exhibit increased sophistication in related domains including Theory-of-Mind (15; 16), pragmatics (17; 18; 19), and modal reasoning (20). From ages four to ten, children also become increasingly aware of the purpose or 'spirit' and scope of a rule. (21; 22; 23) This prior work suggests loophole behavior may emerge around age five, and continue to develop through middle childhood.

Loopholes are pervasive, consequential, and useful for safer goal-comprehension frameworks. Yet, to our knowledge there is no detailed study of how humans learn to find these creative workarounds. Here, we first investigate the emergence and prevalence of loophole behavior in naturalistic settings via a parent survey (Study 1). We then present an initial experiment on children's and adults' understanding of loopholes (Study 2). 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 human-AI interaction.

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

We surveyed 260 parents online via Prolific about their own children's engagement with loopholes (N = 425 3- to 18-year-olds; M_{age}: 8.7). Participants were given a definition of loophole behavior and 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.

Survey responses indicate that loophole behavior (1) is easily recognized by parents: 93% correctly identified it and many recalled specific instances of such behavior in their own children; (2) is prevalent and frequent in parent-child interactions: 60% were reported as engaging in loophole behavior currently (45%) or previously (15%); (3) emerges around 5 to 6 years (M_{age}: 5.6, range: 2 to 13 yrs), peaks around 7 to 8 (M_{age}: 7.4, range: 2 to 13 yrs), and tapers off around 9 to 10 (M_{age}: 9.3, range: 3 to 17 yrs); (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 (see Fig. 1A-B and Appendix).

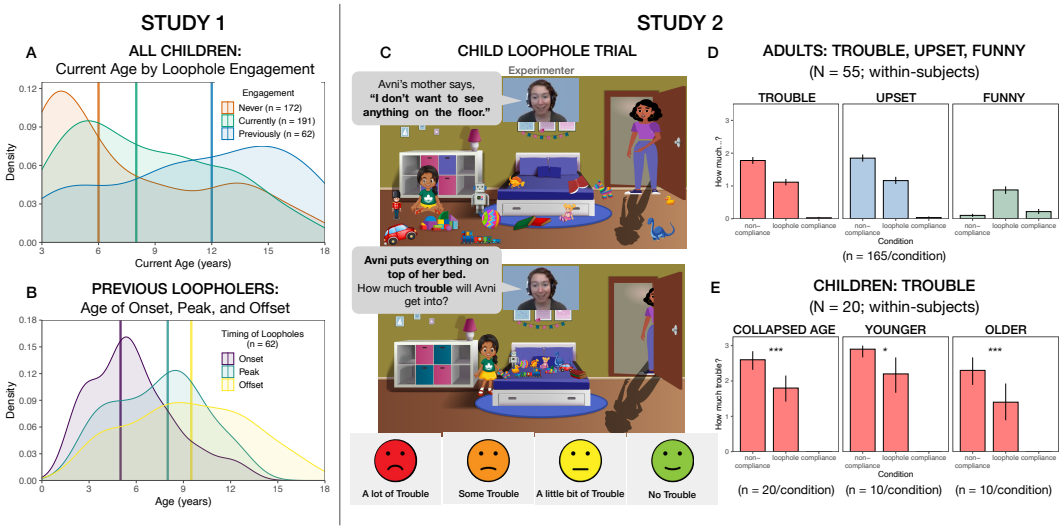

Figure 1: **Study 1: A** Distribution and median age of children reported to have engaged never (red), currently (green), or previously (blue) in loopholes. **B** Distribution and median age of loophole onset (purple), peak (turquoise), and offset (yellow) for previous loopholers. **Study 2: C** Example loophole scenario and trouble scale for children. **D** 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). **F** Children's ratings of trouble: Collapsed Age, Younger, and Older (median age split). Error bars: 95% bootstrapped CIs

3 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, but what do children understand about loopholes? Loophole-behavior may serve to achieve one's own goals, while reducing the severity of retribution, compared to outright non-compliance. This reduction in severity can be due to feigned confusion, perceived cleverness, or even amusement, especially among parents and children. In Study 2, we empirically tested whether adults and children estimate that loopholes decrease the likely degree of punishment and parental upset, as well as increase likely amusement compared to non-compliance.

3.1 Participants and Design

Adult Participants were recruited online via Prolific (N = 55; M_{age}: 32.5, range: 18 to 65 yrs). 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), tested online over Zoom, and thus we consider this study a pilot experiment. All participants were presented with scenarios (based on the real-life examples provided in Study 1) in which a parent made a request of a child, and the child either complied, did not comply, or found a loophole. Adult participants read nine scenarios (3 compliance, 3 non-compliance, 3 loophole) in a Qualtrics survey and evaluated the child's response on a 4-point scale according to (1) how much trouble the child would get into, (2) how upset the parent would be, and (3) how funny the parent would find the behavior. Children saw three scenarios (1 loophole, 1 compliance, 1 non-compliance), presented as illustrated stories, and only evaluated the behavior in terms of trouble. As an exploratory measure, we also coded children's smiles and laughter to see if they found a behavior amusing. (See Fig 1C and Appendix.)

3.2 Results

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 ($\beta = 0.77, SE = 0.09, t = 8.54$). Similar to adults, both older and younger children (median-age split) 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 ($\beta = 0.96, SE = 0.28, t(14.24) = 3.49, p = .004$). (See Fig. 1E-F and Appendix for more details.) Children also smiled and/or laughed more for loopholes (40%) than compliance (5%) or non-compliance (5%). 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.

4 Discussion

We present two studies that 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 (Study 1). Adults' and children's evaluations of loopholes vs. non-compliance were consistent with the hypothesis that loopholes can be a means to achieve one's own goals, while reducing the severity of social penalty (Study 2). Four- to 10-year-olds thought exploiting a loophole would result in less trouble than non-compliance, and this belief may increase with age. These findings parallel the developmental trajectory of loophole behavior in Study 1, suggesting that children's ability to identify others' loophole behavior may correlate with the degree to which they exploit loopholes themselves, as well as advancements in other related domains (e.g., Theory-of-Mind and pragmatic reasoning). (17; 15)

This work is a first step in a more detailed empirical and formal study of the development of loophole behavior. Parent report is informative, but limited, as it relies both on parents' memory and ability to correctly identify loophole behavior. Children's responses in Study 2 are consistent with the idea that loophole understanding emerges between four and ten years of age, but this is a preliminary study. In order to more robustly and precisely interrogate the developmental trajectory of loophole behavior, we are currently replicating Study 2 with a larger, more diverse sample and more scenarios. We are also exploring when children predict others will exploit loopholes given varying costs of compliance and non-compliance. This work provides the basis for constructing a formal framework.

Finding and exploiting a loophole brings together three separate components: what is being asked (the speaker's goal), what are my own goals (the listener's goal), and how best to align the two (the trade-off between goals). The Rational Speech Act (RSA) framework formalizes communication as a cooperative act between a speaker who attempts to convey the state of the world to a listener who strives to accurately understand them. (24; 25) We propose that intentional misunderstandings could arise within an RSA setup that combines rational planning models with a joint-utility framework.

In a standard **RSA** setup, a speaker and listener collaborate to reason about a space of intended meanings. (24) Given a specific utterance, the listener considers a speaker whose utility is typically linked to whether the listener correctly infers the intended meaning. Our framework will integrate this RSA model with **planning frameworks**, specifically expected utility maximization. (26; 27) In our framework, the intended meaning is itself the speaker's utility (goal). The listener chooses actions to maximize their own utility, while also taking into account that of the speaker, leading to collaborative or helpful acts through **joint planning**. (28; 26) Standard goal-communication pipes the speaker's utility (the output of RSA) into planning to produce an action. A low utility outcome for the listener, however, could 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). (29) 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.

A formal understanding of how humans learn to intentionally find loopholes can help design machines that learn to better understand people and avoid misalignments in human-technology interactions. Engineered systems currently do not have human-like goals, but nonetheless behave, at times, like humans exploiting loopholes. Current efforts to promote AI Safety focus more on engineering safety than understanding how humans reason about goals. (30) To extend our formal framework for loophole behavior from human-human to human-machine communication, we first plan to investigate how people think about machines as social partners, or their 'intuitive theory of AI.' What assumptions

do people hold about machine vs. human abilities to understand goals, and do they spontaneously correct for possible misalignment when instructing machines, but not other humans? In parallel, we will look at how interaction models and interface affordances can help humans interactively provide (still partial) specifications that are more likely to result in behavior they want from (a) a general AI and (b) the particular AI they are interacting with, sans apparent loophole behavior.

Loopholes subvert the usual process of goal inference and joint action. In doing so, they offer a different lens for the typical workings of cooperation and reasoning about intention, just as visual illusions shed light on the implicit assumptions and computations of the visual system. The current work lays a foundation for developmentally and computationally characterizing loopholes, supporting new frameworks for analyzing pragmatic communication and social decision making.

5 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 Vélez for their helpful comments and discussion. This research is funded by a MIT Simons Center for the Social Brain Postdoctoral Fellowship (SB), the MIT Center for Brains, Minds, and Machines (TU), and a NSF Science of Learning and Augmented Intelligence Grant 2118103 (EG, LS, TU).

References

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

[2] Isenbergh, J. Musings on form and substance in taxation (1982).

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

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

[5] Uther, H.-J. *The types of International Folktales–A classification and bibliography* (Suomalainen Tiedeakatemia Academia Scientiarum Fennica Exchange Centre, 2004).

[6] Russell, S. *Human compatible: Artificial intelligence and the problem of control* (Penguin, 2019).

[7] Amodei, D. *et al.* Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565* (2016).

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

[9] Lehman, J. *et al.* The surprising creativity of digital evolution: A collection of anecdotes from the evolutionary computation and artificial life research communities. *Artificial Life* **26**, 274–306 (2020).

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

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

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

[13] Woodward, A. Infants selectively encode the goal object of an actor's reach. *Cognition* **69**, 1–34 (1998).

[14] Jara-Ettinger, J., Gweon, H., Schulz, L. E. & Tenenbaum, J. B. The Naïve Utility Calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences* **20**, 589–604 (2016).

[15] Tomasello, M. How children come to understand false beliefs: A shared intentionality account. *Proceedings of the National Academy of Sciences* **115**, 8491–8498 (2018).

[16] Tomasello, M. The role of roles in uniquely human cognition and sociality. *Journal for the Theory of Social Behaviour* 2–19 (2020).

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

[18] Skordos, D. & Papafragou, A. Children's derivation of scalar implicatures: Alternatives and relevance. *Cognition* **153**, 6–18 (2016).

[19] Demorest, A., Silberstein, L., Gardner, H. & Winner, E. Telling it as it isn't: Children's understanding of figurative language. *British Journal of Developmental Psychology* **1**, 121–134 (1983).

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

[21] Heyman, G. D., Sweet, M. A. & Lee, K. Children's reasoning about lie-telling and truth-telling in politeness contexts. *Social Development* **18**, 728–746 (2009).

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

[23] Bregant, J., Wellbery, I. & Shaw, A. 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 (2019).

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

[25] Frank, M. C. & Goodman, N. D. Predicting pragmatic reasoning in language games. *Science* **336**, 998–998 (2012).

[26] Russell, S. & Norvig, P. *Artificial Intelligence: A Modern Approach* (Pearson, 2020), 4th edn.

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

[28] Ullman, T. D. *et al.* Help or hinder: Bayesian models of social goal inference. *Advances in Neural Information Processing Systems* 1874–1882 (2009).

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

[30] Leike, J. *et al.* Ai safety gridworlds. *arXiv preprint arXiv:1711.09883* (2017).

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [N/A]

(b) Did you mention the license of the assets? [N/A]

(c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] (Please see main text and Appendix.)

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes] (Please see Appendix.)

A Appendix

Here, we provide additional details on the methods and results for Study 1 and Study 2. The code, data, and instructions needed to reproduce the main experimental results in the paper can be found on the Open Science Framework (OSF) at `https://osf.io/rwgmx/`. All recruitment, consent, and study procedures were approved by the Institutional Review Boards of MIT and/or Harvard University. All adult participants and parents of child participants provided informed consent to participate.

B Study 1: Parent Survey

B.1 Participants

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, Latinx, or Spanish Origin, 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). The survey took approximately 9 minutes and compensation was $1.43 (approximately $9.53/hr; total spent on participant compensation: $561.35). All participants read a consent form and indicated their consent to participate by selecting the corresponding radio button.

B.2 Procedure

The definition of loophole behavior, along with the examples, provided to participants, as well as the instructions for the classification trials and the behaviors participants were asked to classify, are shown in Fig. 2. In the classification trials, participants read one story of a child finding a loophole and another of a child refusing to comply and were asked to classify the behavior as either loophole behavior, genuine misunderstanding, or refusing to comply. Participants were given feedback on their

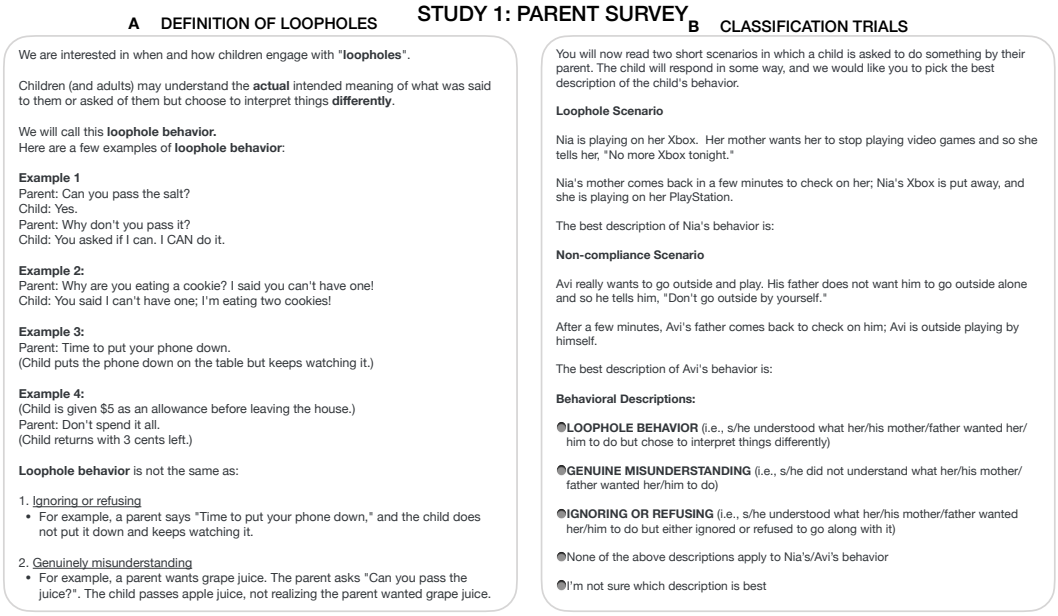

Figure 2: **Study 1: A** Definition of loophole behavior, and examples provided to participants. **B** Instructions for the classification trials and the behaviors participants were asked to classify.

classifications. Participants were then asked if they felt they understood what we meant by loophole behavior and could respond by selecting *Yes*, *No*, or *Maybe*.

Participants were then told that they would be asked about whether or not their own children have ever engaged in loophole behavior. Participants entered the number of children who were between the ages of 3 and 18 years (inclusive), and then for each child, participants were asked the child's age, gender, and whether the child ever engages in loophole behavior. Participants could respond: *Yes, they currently engage in such behavior or recently have*, *They used to engage in such behavior but no longer do*, or *No, they have never engaged in such behavior*. If participants selected *Yes...*, they were asked to report to the best of their recollection: (1) how frequently they would say the child engages in loophole behavior (*several times a day, about once a day, once every few days, once every few weeks, less frequently than once every few weeks*) and (2) at what age the child began engaging in loophole behavior. If participants selected *They used to...*, they were asked at what age the behavior began, when it stopped, when it peaked, and how frequent the behavior was at its peak. Parents of current and previous loopholers were then invited to share a short example (or several examples) of an interaction in which their child engaged in loophole behavior. If participants selected *No...*, they were asked if they had ever observed another child engaging in loophole behavior, and if so, they were invited to share the story and the age of the child if they could recall it. The survey ended with optional demographic questions and a comprehension check (participants were asked what the task was about). If participants entered a nonsensical or irrelevant response (e.g., "Homework") or indicated that they did not know what the task was about, they were excluded from analysis.

B.3 Results

Classification accuracy was high both for loophole behavior (93% correctly identified it) and non-compliance (91%). After the classification trials, 97.7% of participants selected *Yes* when asked if they understood what was meant by loophole behavior, 1.9% selected *Maybe* and .4% selected *No*. In total, parents shared 256 anecdotes of their children's behavior and 206 of these anecdotes (80%) were examples of loophole behavior.

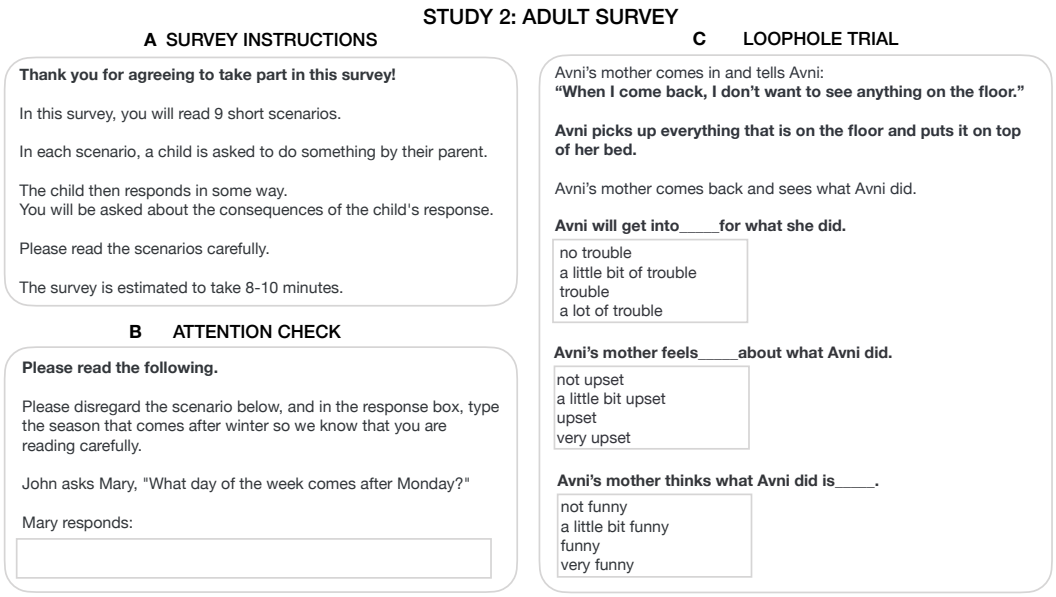

Figure 3: **Study 2: A** Instructions provided to participants. **B** Attention check. **C** Example of a loophole trial and dependent measures.

C Study 2: Experiment with Adults and Children

C.1 Adult Experiment

C.1.1 Participants

Participants (N = 55; M_{age}: 32.5, range: 18 to 65 yrs, 55% female, 42% male, 2% trans male, 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 (approximately $10.73/hr; total spent on participant compensation: $114.41). Participants were majority White (64%; 11% Black, 11% Hispanic, Latinx, or Spanish-Origin, 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 (described below). All participants read a consent form and indicated their consent to participate by selecting the corresponding radio button.

C.1.2 Procedure.

We created 27 different scenarios (9 stories with 3 endings each) based on real-life examples provided in Study 1. The order of the nine scenarios participants read 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. Participants then completed an attention check in which they were told to ignore the scenario below and enter the season that comes after winter in the text box. Participants were only included in analysis if they entered "spring" or "Spring" into the text box.

Next, participants were presented with the actual scenarios of the experiment. 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. See Fig. 3 for more details. The scenarios can be found at: `https://osf.io/rwgmx/`

C.1.3 Results

For the results reported in the main text, 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 random effects structure that converged (random intercepts and effects of condition and measure by subject and scenario).

C.2 Pilot Child Experiment

C.2.1 Participants

As discussed in the paper, the sample of child participants was a convenience sample recruited by word-of-mouth to take part in a study over Zoom. We recruited twenty 4- to 10-year-old children located in the U.S. and the U.K. (M_{age}: 6.7, range: 4.1 to 10.1 yrs, 40% female, all White). Parents were sent a consent form and provided their verbal consent at the start of the testing session both for their child's participation and for the session to be video-recorded. At the end of the session, parents indicated video-sharing permissions. Children received a certificate of participation to thank them for their participation. The experiment took approximately 10 minutes to complete and the entire testing session lasted approximately 20 minutes.

C.2.2 Procedure

We created five scenarios based on the scenarios used in the experiment with adults and collaborated with an artist on Fiverr to illustrate them, yielding 15 scenarios in total (5 stories with 3 endings each). Children saw three scenarios (1 compliance, 1 loophole, 1 non-compliance) displayed and narrated over Zoom by the experimenter. The three scenarios children saw, the condition of each scenario and the order of the conditions were pseudo-randomized 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, after hearing the parent's request, children were asked to repeat it to encourage them to pay attention; the experimenter then re-stated the request, so children heard the request twice and stated it once. After learning how the child protagonist responded to the request, children were asked: "How much trouble will (child protagonist) get into for (behavior)?" Children indicated the level of trouble on a 4-point scale, with each point represented as a different colored face expressing a different affect. Children received training and practiced using the scale ahead of time. Children could indicate their choice by the label of the face, the color of the face, or both. Finally, for each scenario, children were asked, "And why will (child protagonist) get into (selected level of trouble)?"

D Results

We conducted a mixed effects linear regression predicting children's ratings of the degree of trouble on a 4-point scale (number 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. A few children said the child protagonist would get into an amount of trouble in between two points on the scale; for these responses, which were rare (n = 4 / 60), we coded them at the mid-point between the integers that corresponded to the two scale-points (i.e., 'in between a little bit of trouble and some trouble' would be coded as 1.5). All other responses were coded as an integer from 0 - 3. The younger age group (4.1 - 6.1 yrs) consisted of children who were below the median age (6.2 yrs), and the older age group (6.2 - 10.1 yrs) consisted of children equal in age to or older than the median age.