Skirting the Sacred: Moral Violations Make Intentional Misunderstandings Worse

Kiera Parece[1, 2], Sophie Bridgers[1, 2], Laura Schulz[1], Tomer D. Ullman[2]

[1] Department of Brain and Cognitive Sciences, MIT
[2] Department of Psychology, Harvard University

Abstract

People engage in intentional misunderstandings to get around direct non-compliance. In other words, they use loopholes. Previous work showed that adults and children consider loophole behavior to be less costly than direct non-compliance (Bridgers, Schulz, & Ullman, 2021), and suggested this is a primary reason for their use: loopholes will land you in less trouble than defiance. However, we propose that this difference between loopholes and defiance will not hold for a specific, important context: moral violations. We replicate the finding that loopholes are less costly in a neutral context but find that engaging in loopholes in a moral context is as bad as non-compliance (Experiment 1, N=360). We then use a direct comparison between loopholes and non-compliance (N=150) to investigate whether in some contexts loopholes will be seen as even worse than non-compliance. We replicate the differential effect of the moral context from Experiment 1, but do not find a reversal. We discuss possible extensions and limitations, and consider why loopholes in moral violations may be uniquely unacceptable.

Keywords: loopholes; goal alignment; communication; morality

A teenager sits on the sofa, listening to loud music. Their mother walks by and snaps: "put some headphones on!". The teenager pulls out headphones, places them on their head, and continues listening to their blaring music out loud. The teenager understood what their mother wanted but acted on an alternative interpretation.

When a person uses the ambiguity of language by intentionally misunderstanding a request, they are engaging in 'loophole behavior'. This behavior is ubiquitous: history, literature, law, and everyday life are full of examples of people who do what they're asked, but not what they're told (Isenbergh, 1992; Fuller, 1957; Katz, 2010; Scott, 1985; Uther, 2004). Even young children use loopholes in day-to-day life, when confronted with situations in which their goals are at odds with another person. By engaging in a loophole, people pursue their own goals, while reducing the repercussions of outright noncompliance.

Loopholes are also an increasingly pressing issue in the fields of Artificial Intelligence and Machine Learning, where such behavior may go under different terms, such as 'specification gaming', 'malicious compliance', and 'cheating' (Amodei et al., 2016; Russell, 2020; Everitt et al., 2021). Machines that do what they're asked but not what we want pose safety risks and have become a major focus of both researchers and policy makers (Amodei et al., 2016).

Recent empirical work has also examined how people evaluate loopholes in interactions between parents and children and found that loopholes are perceived as a method to achieve one's own goals, while reducing a possible penalty. Both children and adults thought a child who exploited loophole behavior would get into less trouble with their parent than a child who directly defied a request. They also thought loopholes were funnier than either compliance or non-compliance (Bridgers, Schulz, & Ullman, 2021).

While the previous findings indicate loopholes are less costly than outright non-compliance, is this *always* the case? The question is important as loopholes seem to primarily rely on getting one into less trouble. But intuition suggests that in some situations they may be as bad as or worse than non-compliance. Consider a man who recently discovered his wife is having an affair. Confronting his wife, the man tells her, "I want you to stop seeing Bill." His wife proceeds to end the affair with Bill, only to start an affair with Ted. While the behavior of the surly teenager may be exasperating, that of the hairsplitting adulteress seems even more outrageous than a stark refusal. The basic behavior already violates fidelity, the loophole treats it as a joke.

The examples of the loud teenager and lawyerly partner vary in many particulars, but we propose that the main distinction that may make a difference is the existence or absence of a moral concern. We suggest that loopholes are not always less costly nor more amusing than non-compliance. In particular, we predict that exploiting a loophole that violates a moral principle will result in equal, or perhaps even greater, penalties than outright non-compliance, and will not be considered funny. In the rest of the introduction, we consider some research from psychology and law which motivates the focus on the moral domain in particular, and briefly summarize our own findings.

Hannikainen et al. (2022) recently examined the discrepancy between violating the letter of a law (the literal meaning) versus the spirit of a law (the purpose or intent of the law) by surveying how people evaluate these different types of violations. Participants were presented with a series of vignettes where either the letter or the spirit of a law was broken, and were then asked in each scenario to assess how morally blameworthy the violation was, and whether the law had been violated. Results from this study showed that people consider acts that violate the spirit of a law, but not the letter (as loopholes do) to be morally blameworthy, while acts that breach the letter of the law but not the spirit not to be morally blameworthy. In contrast, when participants were asked which act violated the law itself, people concluded that acts that break the spirit (but not the letter) of the law are in accordance with the law, while acts that violate the letter (but

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

not the spirit) of the law breach the law. These findings demonstrate that adults place greater weight on the spirit or purpose of a rule (compared to the rule's literal meaning) when gauging moral culpability, and establish that adults construe intentional violations of the spirit of a rule as grounds for moral blameworthiness.

Garcia, Chen, & Gordon (2014) conducted a series of experiments investigating the same dichotomy between spirit and letter of the law in legal contexts, where people were asked to evaluate the culpability of a protagonist (either themselves or a third-party actor) for breaking either or both the spirit and letter of the law. Similar to the findings of Hannikainen et al. (2022), these results showed that people do not ascribe culpability based on the letter of the law, and tend to assign culpability when the spirit of the law has been violated. This study also showed that culpability judgments still occur even when the letter of a law remains intact, and only the spirit is broken. Taken together, these two studies suggest that loopholes, which violate the spirit of an utterance while maintaining the letter, are at times seen as acts worthy of moral blame.

Bregant et al. (2019) explored how the spirit-letter duality of rules influences children's moral evaluations. Similar to Garcia et al. (2014), they examined the inverse of loopholes, specifically whether children consider the act of violating the letter of a rule, while leaving the spirit of the rule unbroken, to be less wrong than breaking both the letter and spirit. The results showed that from ages four to nine, children become more lenient in their evaluations of actions that violate the letter of a rule yet keep the rule's spirit intact.

All of the findings discussed so far lead us to believe that loopholes, which break the spirit of a rule and not the letter, can warrant moral blame from both adults and children. The question remains however whether this blame will be less than, equal to, or greater than outright refusal. The previous studies on letter-vs-spirit of the law have not considered that a primary *use* of loopholes in a social context is to get around requests, and they are often useful to the degree that they incur less cost, whereas the previous study of loopholes in a social context (Bridgers et al., 2021) did not consider moral violations.

What may drive the difference between a neutral context and a moral context? We expand on this question at the end of the paper after laying out the empirical findings, but even prior to empirical data, we can point to the fact that people's moral judgments and evaluations of culpability appear to rely on a dual-process framework: one process that considers the intent of an agent, and another that weighs the outcome of the act itself or the harm caused (Cushman, 2008; Young, Cushman, Hauser, & Saxe, 2007). In neutral scenarios, ambiguity surrounding the intent of a speaker is more believable under the pretense of plausible deniability, where a listener could justify their actions in the space of alternative interpretations of the utterance. In moral scenarios, while this ambiguity of intent remains, the listener's response or action may be inherently problematic, consisting of harm or malice, and creating situations where the speaker's technical request becomes less important or relevant in light of the problematic behavior of the loophole itself. In terms of outcome in the dual-process framework, employing a loophole in both neutral and moral contexts fails to achieve the desired outcome or the goal of the speaker. However, one crucial distinction between these two contexts remains: one results in the causation of harm.

To summarize the main argument in brief then: Loopholes are important and pervasive in daily life, and are increasingly a topic of study in law, cognitive science, and AI/ML. One of the primary uses of loopholes is to get around requests that conflict with one's own goals. The use of loopholes as dodges is valuable to the degree that loopholes result in less cost, trouble, or harm compared to non-compliance. However, there is reason to think that loopholes that are moral violations will be judged just as bad, or even worse than outright refusal or overtly asocial behavior (i.e. you don't joke about some things). If this turns out to be empirically true, it would inform us both about the structure of loopholes, and about moral reasoning as a separate domain within social reasoning.

In two experiments, we tested the proposal that loopholes in moral situations are not acceptable, and will not be less costly than non-compliance. In both studies, participants read vignettes in which a speaker made a request of a listener. Vignettes either described a "neutral" context (one in which a loophole or defiance is not a moral violation), or a "moral" context. In Study 1, the listener responded by either complying, not complying, or exploiting a loophole in the request. Participants were asked to evaluate the behavior in terms of how much trouble the behavior would incur, how upset the speaker would be, and how amused the speaker would be. In Study 2, participants were asked to directly compare non-compliant and loophole responses, and select which behavior would get the listener into more trouble with the speaker.

As we describe in more detail below, we found that in neutral contexts, loopholes are seen as likely to lead to less trouble and upset than non-compliance, and were also more amusing. These results replicated the previous findings of Bridgers et al. (2021). Importantly, however, when moral transgressions occurred, loopholes were seen as being as bad as non-compliance, though they were not seen as *worse* than non-compliance. After detailing the experiments and findings, we discuss the implications of the work, consider the limitations of the current studies, propose future avenues of research, and speculate on why the moral context matters for loopholes.

Study 1: Comparing loopholes in neutral and moral contexts

We first examined whether people evaluated loopholes differently compared to compliance and non-compliance, when the loophole was a neutral transgression vs. a moral transgression. We predicted that when the loophole was a neutral behavior, we would replicate prior findings and

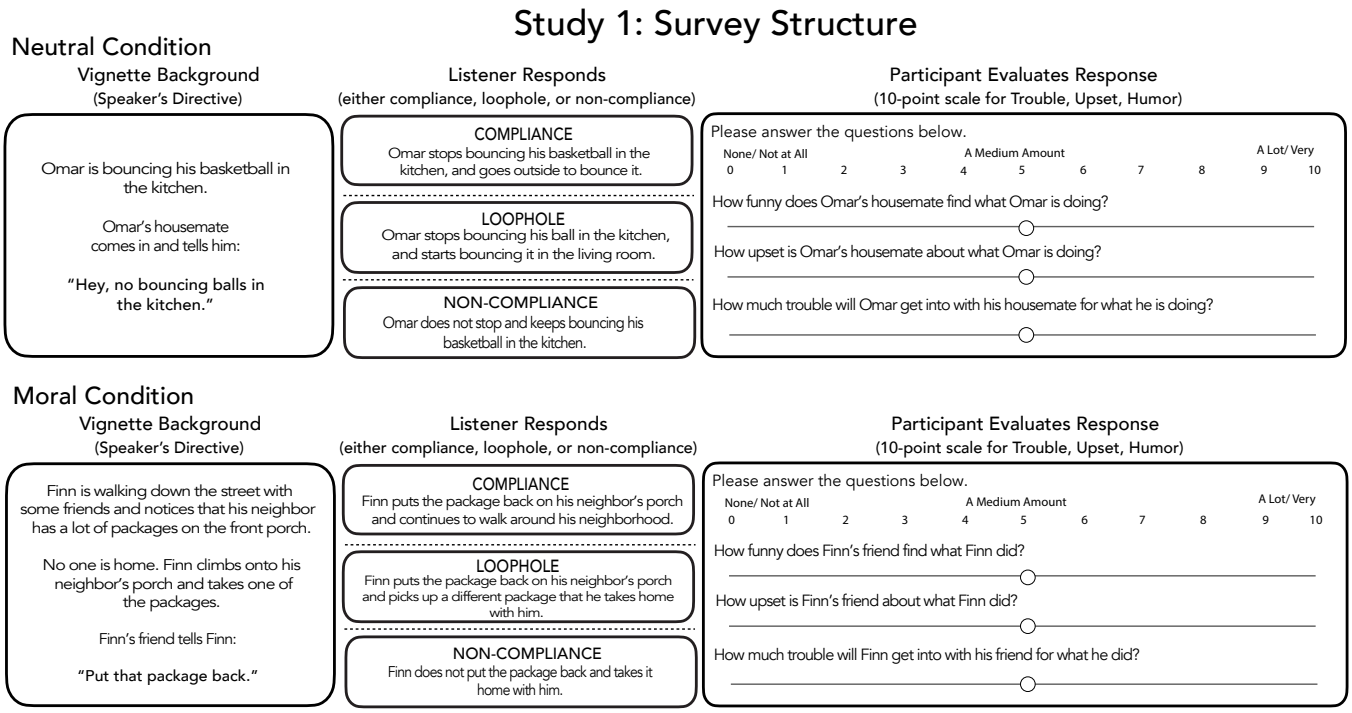

Figure 1: *Survey Structure for Study 1.* Two between-subjects surveys were used to assess how people evaluate loophole behavior, compared to compliance and noncompliance. Participants either completed the neutral condition, or the moral condition. Each condition included vignettes that featured a speaker and a listener, in which a speaker makes a request and a listener responds with one of 3 types of behavior (compliance, loophole, non-compliance). People then assessed the behavior in terms of how much trouble it will lead to, how upset the speaker will be, and how amused the speaker will be.

participants would rate it as less problematic than non-compliance (though more than compliance), and that it would be deemed more humorous than either non-compliance or compliance. However, when the loophole was a moral transgression we predicted it would be rated as equally, or even more problematic than non-compliance, and would not be considered funny.

Participants. We recruited 360 adults with above a 95% approval rating online via Prolific (see Peer et al., 2017). Participants (M_{age}: 39.28; 46.94% female, 2.5% nonbinary; 73.89% White, 6.94% Hispanic or Latinx, 6.11% Black/ African American, 5.55% Mixed, 5% Asian, 2.22% Other) were U.S. residents fluent in English and from a range of geographical regions and educational backgrounds. An additional 37 participants were recruited but excluded from analysis due to either a failure to pass an attention check (N = 30) or already having participated in a closely related study (N = 7).

Participants were randomly assigned to one of two between-subjects conditions: the Moral (n = 180; 47.77% female, 2.22% nonbinary) and the Neutral condition (n = 180; 46.11% female, 2.77% nonbinary). The surveys took approximately 9 minutes to complete, and compensation was $2.14.

Procedure. We created two separate Qualtrics surveys corresponding to the Neutral and Moral conditions (see Figure 1). In each survey, participants read 12 short scenarios in which one person (a speaker) made a request or directive to another person (a listener). In each scenario, the listener could respond in one of three possible ways to the speaker's request: (1) comply with the request, (2) directly not comply with the request, or (3) find a loophole in the request (technically fulfilling the speaker's request but capitalizing on ambiguity to act on an unintended meaning). Participants only saw one version of each of the 12 stories (i.e., one response to the speaker per story). For each vignette, participants were asked to rate on a 10-point numeric scale (1) how much trouble the listener would get into for the given behavior, (2) how upset the speaker would be about the listener's given behavior, and (3) how funny the speaker would find the listener's behavior (with "0" signifying None / Not at all and "10" representing A Lot / Very).

In all scenarios, the speaker and listener were both adults, which was made explicit to participants. Each scenario also had a power relationship between the speaker and listener such that the listener could be in a position of equal or lesser power to the speaker (e.g., the speaker and listener are co-workers, or the speaker is the listener's boss). Power relationships between the speaker and listener were fixed within scenario, but balanced so that 6 stories were up-relationships and six were equal; the types of relationships (e.g. professional or familial) were also balanced across surveys. Stories were randomized with the constraint that

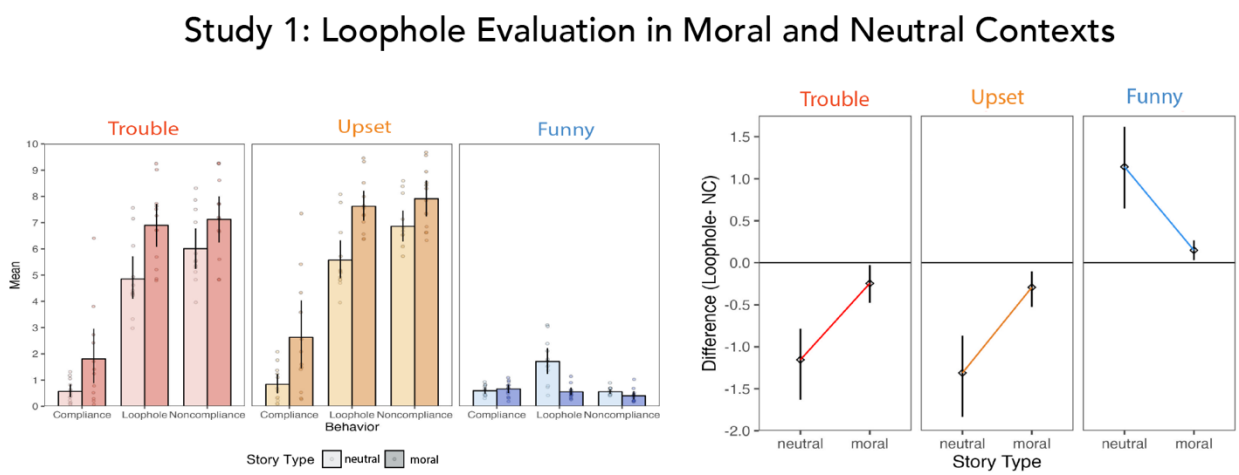

Figure 2: *Results of Study 1.* On the left, mean ratings of trouble (red), upset (orange), and humor (blue) are shown by behavior type and condition (story type); points are scenario (item) means. The graph on the right depicts the mean difference calculated between loophole behavior and non-compliance between story type by measure. Error bars are bootstrapped 95% Confidence Intervals (CIs).

each participant read 4 scenarios where the listener responded with compliance, 4 with a loophole, and 4 with non-compliance. The survey for the Neutral condition consisted only of scenarios where protagonists engaged in loopholes in neutral contexts, while the survey for the Moral condition consisted exclusively of scenarios where protagonists engaged in loopholes in moral contexts.

In the Moral condition, the scenarios were developed using the framework and core tenets of Moral Foundations Theory: harm, cheating, betrayal, subversion, and degradation (Graham et al., 2013). In the Neutral condition, scenarios were inspired by real-life anecdotes collected through ongoing work.

Results. We conducted three mixed effects linear regressions, one for each dependent measure, predicting participants' ratings of the amount of trouble, upset, and funniness that a behavior garnered from fixed effects of condition (story type) and behavior, and their interaction, with random intercepts and effects of behavior by subject and by story/scenario. Participants' responses were a numeric value between 0-10, condition is a 2-level between-subjects factor (moral vs. neutral), and behavior is a 3-level within-subjects factor (compliance, loophole, non-compliance).

In the Neutral condition, participants rated loopholes as resulting in more trouble and upset than compliance (trouble: $\beta = -4.299$, $SE = 0.471$, $t(23.614) = -9.137$, $p < .001$, upset: $\beta = -4.758$, $SE = 0.493$, $t(23.569) = -9.635$, $p < .001$), but less trouble and upset than direct non-compliance (trouble: $\beta = 1.143$, $SE = 0.182$, $t(24.282) = 6.286$, $p < .001$, upset: $\beta = 1.279$, $SE = 0.190$, $t(22.248) = 6.727$, $p < .001$). Participants also rated loopholes higher in terms of funniness compared to compliance ($\beta = -1.106$, $SE = 0.178$, $t(24.157) = -6.217$, $p < .001$) and non-compliance ($\beta = -1.132$, $SE = 0.175$, $t(22.526) = -6.449$, $p < .001$). In the Moral condition, participants again rated loopholes as resulting in a higher degree of trouble and upset than compliance (trouble: $\beta = -5.058$, $SE = 0.471$, $t(23.613) = -10.746$, $p < .001$, upset: $\beta = -5.027$, $SE = 0.494$, $t(23.583) = -10.175$, $p < .001$), however, provided equal ratings of trouble and upset compared to non-compliance (trouble: $\beta = 0.250$, $SE = 0.182$, $t(23.953) = 1.377$, $p = .181$; upset: $\beta = 0.304$, $SE = 0.190$, $t(22.036) = 1.599$, $p = .124$). Participants in the Moral condition also rated loopholes as generating the same amount of humor as compliance ($\beta = 0.130$, $SE = 0.178$, $t(24.056) = 0.732$, $p = .471$) and non-compliance ($\beta = -0.149$, $SE = 0.175$, $t(22.399) = -0.849$, $p = .405$), suggesting that loopholes in moral contexts are not funny.

Across conditions, loopholes were rated as incurring more trouble ($\beta = -2.010$, $SE = 0.612$, $t(25.326) = -3.282$, $p = .003$), and upset ($\beta = -2.036$, $SE = 0.496$, $t(26.052) = -4.104$, $p < .001$), and as less funny ($\beta = 1.142$, $SE = 0.271$, $t(31.218) = 4.221$, $p < .001$), in the moral compared to the neutral condition. There were also significant interactions between condition and behavior. In particular, the difference between loopholes and non-compliance was greater in the neutral condition compared to the moral condition in terms of trouble ($\beta = 0.894$, $SE = 0.257$, $t(24.116) = 3.478$, $p = .002$), upset ($\beta = 0.976$, $SE = 0.269$, $t(22.140) = 3.631$, $p = .001$), and humor ($\beta = -0.984$, $SE = 0.248$, $t(22.461) = -3.964$, $p < .001$). (See Figure 2.)

In summary, in the Neutral vignettes people thought loopholes were not as bad as direct non-compliance. In the moral condition, there is an overall effect such that both loopholes and non-compliance are seen as worse than their counterpart in the neutral condition. But crucially, loopholes were differentially affected by the moral vignettes, as can be seen by the shrinking distance between loopholes and non-compliance in Figure 2 (right). A similar dynamic was found for ratings of amusement, confirming that people think loopholes in the moral domain are much less amusing than in more neutral contexts.

Study 2: Direct comparison of loopholes and noncompliance

In Study 1, we established that loopholes do not always reduce costs, and that when a loophole violates a moral principle, the expected penalty is similar to that of outright noncompliance. As participants in Study 1 viewed and evaluated only one behavior per story (either compliance, loophole, or noncompliance), it remained an open possibility that a direct comparison between two types of behavior would yield a different pattern of evaluations. In particular, we were interested to see whether such a direct comparison would result in a *reversal*, such that loopholes may be seen as even worse than non-compliance, in a moral context. So, in Study 2, we pit loophole behavior against non-compliance behavior directly, and asked adults to assess which behavior would incur more trouble.

Participants. We recruited 150 adults with above a 95% approval rating online via Prolific (Peer et al., 2017). Participants (M_{age}: 32.64; 49.33% female, 1.33% nonbinary; 64% White, 8.67% Hispanic or Latinx, 8.67% Black/ African American, 5.33% Mixed, 9.33% Asian, 4% Other) were U.S. residents fluent in English and from diverse regional and educational backgrounds. An additional 18 participants were recruited but excluded from analysis due to failing an attention check. The survey took 9 minutes to complete, and compensation was $2.14.

Procedure. We used the same Qualtrics surveys corresponding to the Moral and Neutral conditions developed for Study 1 in which one person (a speaker) made a request of another person (a listener). Participants were then presented with two possible behaviors that the listener could choose in response to the speaker's directive (either a loophole behavior or noncompliance behavior). Participants were asked to select which behavior of the two would get the listener into more trouble with the speaker. The order of the scenarios and behaviors were counterbalanced. Whether the loophole behavior was presented first or the non-compliance behavior was presented first was evenly and randomly distributed between the vignettes across participants.

Results. Figure 3 summarizes the findings of Study 2. We conducted a logistic regression predicting participants' choices from a fixed effect of condition. Responses were coded as an integer (either 0 for selecting noncompliance or 1 for loophole) and condition is a 2-level between-subjects factor (moral vs. neutral). This model revealed that participants were more likely to select the loophole behavior as worse in the Moral condition compared to the Neutral condition ($\beta = 0.951$, $SE = 0.116$, $z = 8.157$, $p < .001$). Within both conditions, however, participants were significantly more likely to select the non-compliant behavior as worse than the loophole behavior (Neutral: $\beta = -1.700$, $SE = 0.092$, $z = -18.432$, $p < .001$; Moral: $\beta = -0.748$, $SE = 0.071$, $z = -10.487$, $p < .001$). In other words, while we replicated the findings from Study 1 that (1) in neutral contexts loopholes are seen as less bad than non-compliance and (2) that moral violations had a differential effect on how bad loopholes are considered to be, we did *not* find a reversal of loopholes being even worse than non-compliance.

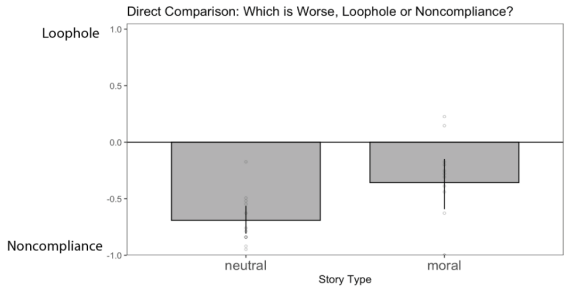

Figure 3: *Results of Study 2.* People were given vignettes and possible actions that corresponded to loophole vs. non-compliance, and asked to directly judge which of the two behaviors was worse. People overall choose non-compliance as worse in this direct comparison, but the difference shrank in the Moral condition compared to the Neutral condition. Graph shows mean choice of loophole (+1) vs. non-compliance (-1) by condition (story type). Points are scenario (item) means and error bars are bootstrapped 95% CIs.

Discussion

Loopholes are common, probably because they are useful. A chief use of loopholes is in dodging requests. In line with this, previous research has shown that loopholes are expected to incur less costs than non-compliance (Bridgers et al., 2021). However, it

appears that there are specific, important situations in which this overall pattern begins to break down: situations that involve moral violations. Some things are not to be toyed with, skirted, feigned, or dodged.

In two experiments, we replicated the previous pattern of loopholes in neutral contexts incurring less trouble than non-compliance (though more than compliance) and resulting in more humor than either non-compliance or compliance (Bridgers et al., 2021). But, we also found that when loopholes violate moral principles, they are expected to result in a similar amount of cost (trouble, upset) as noncompliance, and were no longer regarded as humorous or amusing. These findings expand upon recent work that contends loopholes serve as a tool for reducing costs, and demonstrate that not all loopholes provide a means to skirt social consequences.

The work presented here is only a partial extension of ongoing research examining loopholes. There are several immediate follow-ups one could consider to this work, some of which we are actively pursuing. For example, our experiments, analyses, and results relied on a binary, experimenter-crafted distinction between neutral and moral contexts. People's intuitions about whether something constituted a moral situation, context, or transgression may differ from the experimenters', and the distinction between moral and neutral transgressions is likely to be more continuous than binary in nature. This is not necessarily a challenge to the current findings but a potentially exciting extension. One could ask a group of people to judge the degree to which a scenario presents a moral transgression, and another group of people to judge how bad a loophole is in such a case. We would predict a quantitative effect beyond the simple binary results here, such that the more morally transgressive a situation is seen to be, the worse the use of a loophole in such a case would be.

Another extension to the work could expand the analysis to the sub-types of moral domains under consideration, and individual differences along these lines. The moral principles or transgressions developed in these studies were designed to encompass the multiple facets outlined in Moral Foundations Theory (Graham et al., 2013), but the analysis did not consider these foundations separately. For example, it is possible that the degree to which one rejects or accepts loopholes to get around religious injunctions depends on one's view on sanctity as a moral foundation. As a cautionary aside to this example we would point out that in history religious sub-groups that saw themselves as equally devout still could develop quite different ideas about whether it is legitimate to take a legalistic approach to holy scripture (e.g. the split between *Hasidim* and *Misnagdim*).

Other extensions are possible, but we turn instead to a bigger question unanswered by the current set of studies: Why? Why are loopholes as bad as non-compliance in the moral domain? Our results suggest that they are, as a brute fact, but they do not in themselves explain why this is so.

It is easy to come up with plausible just-so explanations for the findings, but many of them turn out to be circular. For example, perhaps loopholes incur less cost in a neutral context because the listener can fall back on plausible deniability, which is not available in the moral context. Such an explanation may seem reasonable at first but fails on two counts. First, it seems unlikely that the teenager from the opening example could honestly claim to have been confused about what their parent meant. Claims of genuine ignorance or misunderstanding in the scenarios we considered don't pass the giggle test. But more importantly, *why* would plausible deniability be available to the neutral transgressor and not the moral transgressor? The answer would be a version of "well, *obviously* no one would misunderstand a moral request so you can't claim to be ignorant", but this explanation amounts to saying the moral domain is special because the moral domain is special. Similar circularity exists for explanations that are long-form versions of "there are some things you don't joke about". Ultimately, it may turn out that the explanation has to do with the fact that in the moral domain loopholes involve a differential degree of harm, but this remains to be examined.

Before closing, we note that we were not surprised to find that loopholes were differentially affected by the moral context compared to non-compliance, but we were surprised that we did not find a *reversal*, such that loopholes were seen as even worse than non-compliance. It may be that non-compliance places a ceiling on how bad things are taken to be, or it may be that our exploration was too limited.

The unwritten rules of paper-writing say that a paper should include a meaningful final paragraph, preferably one that ties the paper together, and makes an important last point. But there's nothing sacred about that.

Acknowledgements

We thank the individuals who participated in this research, and the members of the MIT Early Childhood Cognition Lab and members of the Harvard Computation, Cognition, and Development Lab for their helpful comments and discussion. This research was funded by an NSF Science of Learning and Augmented Intelligence Grant 2118103 (LS, 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*.

Bridgers, S., Schulz, L., & Ullman, T. (2021). Loopholes, a Window into Value Alignment and the Learning of Meaning. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 43.

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.

Everitt, T., Hutter, M., Kumar, R., & Krakovna, V. (2021). Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective. *Synthese*, *198*(Suppl 27), 6435-6467.

Fuller, L. L. (1957). Positivism and fidelity to law--A reply to Professor Hart. *Harv. L. Rev.*, *71*, 630.

Garcia, S. M., Chen, P., & Gordon, M. T. (2014). The letter versus the spirit of the law: A lay perspective on culpability. *Judgment & Decision Making*, *9*(5).

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology* (Vol. 47, pp. 55-130). Academic Press.

Hannikainen, I.R., Tobia, K.P., de Almeida, G.d.F., Struchiner, N.,Kneer, M., Bystranowski, P., Dranseika, V., Strohmaier, N., Bensinger, S., Dolinina, K., et al. (2022). Coordination and expertise foster legal textualism. *Proceedings of the National Academy of Sciences*, *119*(44), e2206531119.

Isenbergh, J. (1982). Musings on Form and Substance in Taxation.

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

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153-163.

Russell, S. J. (2021). Human-compatible artificial intelligence. In *Human-like machine intelligence.*

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