Understanding rule enforcement using drift diffusion models

Neele Engelmann (engelmann@mpib-berlin.mpg.de)
Center for Humans and Machines, Max Planck Institute for Human Development, Berlin
Ivar R. Hannikainen[1] **(ivar@ugr.es)**
Department of Philosophy, University of Granada
Carlos González-García (cgonzalez@ugr.es)
Department of Experimental Psychology, University of Granada
María Ruz (mruz@ugr.es)
Department of Experimental Psychology, University of Granada

Abstract

Since their inception, drift diffusion models have been applied across a wide range of disciplines within psychology to uncover the mental processes that underlie perception, attention, and cognitive control. Our studies contribute to ongoing efforts to extend these models to abstract, social reasoning processes like moral or legal judgment. We presented participants with a set of social rules, while manipulating whether various behaviors violated the rule's letter and/or its purpose——two independent standards by which to decide what constitutes a transgression. In this framework, cases that violate or comply with both a rule's text and its purpose can be seen as congruent or 'easy' cases, and cases that elicit opposing verdicts as incongruent or 'hard' cases——in a manner analogous to widely-studied conflict tasks in cognitive psychology. We recorded 34,573 decisions made by 364 participants under soft time pressure, and investigated whether hierarchical drift diffusion modeling could explain various behavioral patterns in our data. This approach yielded three key insights: (1) judgments of conviction were faster than judgments of acquittal owing to an overall bias (z parameter) toward conviction; (2) incongruent cases produced longer reaction times than congruent cases (an interference effect), due to differences in the rate of evidence accumulation (v parameter) across case-types; and (3) increases in the ratio of congruent-to-incongruent cases amplified the interference effect on reaction times, by fostering greater response caution—revealed by a larger threshold (or a parameter). Thus, our studies document dissociable effects of the drift diffusion components on rule-based decision-making, and illustrate how the cognitive processes that subserve abstract and social decision-making tasks, such as the enforcement of communal and legal rules, may be illuminated through the drift diffusion framework.

Keywords: cognitive control, conflict task, statutory interpretation, legal reasoning, drift diffusion modeling

The tension between norm adherence and moral virtue is a recurring theme in moral philosophy. Though abiding by the maxim "do not lie" arguably serves us well in most contexts (Sunstein, 2005), there may be circumstances under which lying is permissible or even obligatory (Engelmann, 2023). A broader literature has now documented the tendency for people to perceive certain rule violations as morally acceptable (Awad et al., 2022; Kwon, Zhi-Xuan, Tenenbaum, & Levine, 2023; Kwon, Tenenbaum, & Levine, 2022), and certain instances of seeming compliance i.e., 'loopholes', as nevertheless undermining the rule's deeper spirit (Bridgers, Taliaferro, Parece, Schulz, & Ullman, 2023). This body of evidence can be fruitfully understood by considering the instrumental dimension of rules—that, as well as having a specific formulation or *text* (e.g., "No shoes in the house"), rules are generally intended to serve a legislative goal or *purpose* (i.e., of keeping the floors clean). In the legal domain, where the question whether a person's conduct was in violation of the law or not can have grave consequences, this division between a rule's letter and its spirit has given rise to rival theories of interpretation—often referred to as textualism (Schauer, 1991) and purposivism (Fuller, 1957) respectively—that remain hotly contested today.

A number of empirical reports have turned attention to the way in which people apply written rules, providing evidence that people prioritize a rule's text (Struchiner, Hannikainen, & de Almeida, 2020), but also attend to the rule's purpose, especially when the lawmaker's goal is seen as morally good (Flanagan, Almeida, Struchiner, & Hannikainen, 2023). This pattern reflects the fact that people simultaneously consider the rule's literal formulation, but also their broader *moral attitude* toward the case at hand, and consequently waver when tasked with applying rules to 'hard' cases, in which a behavior complies with the rule's purpose despite violating its text or *vice versa* (Almeida, Struchiner, & Hannikainen, 2023). In these circumstances, the opportunity to reflect appears to bolster adherence to the rule's text (Flanagan et al., 2023), perhaps in recognition that text provides a focal point (Schelling, 1960) that facilitates coordinated interpretation (Hannikainen et al., 2022).

Overall, these studies provide convergent evidence that people are divided between applying textualist and purposivist standards, and that they subjectively recognize this tension. These findings motivated the objectives of our present research: to examine the extent to which cognitive conflict arises in rule enforcement by exploiting an analogy with basic interference tasks in cognitive psychology and leveraging the toolkit of cognitive modeling to help characterize the psychological processes that subserve rule-based reasoning.

Interference tasks, such as the Stroop and Flanker tests, examine people's ability to classify target stimuli along a task-relevant dimension, while simultaneously manipulating a task-irrelevant dimension between the congruent and incongruent conditions. Repeatedly, researchers have documented longer reaction times and reduced accuracy in incongruent trials relative to congruent trials—a phenomenon known as the *interference effect*—and efforts to apply drift diffusion mod-

[1]Corresponding author

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

eling (Ratcliff, Smith, Brown, & McKoon, 2016; Ratcliff, 1978; Ratcliff & Smith, 2004; Ratcliff & McKoon, 2008; Myers, Interian, & Moustafa, 2022; Voss, Rothermund, & Voss, 2004) have yielded a deeper understanding of the cognitive processes that underlie this effect. Taking binary responses and response times as input, drift diffusion models estimate four main parameters which jointly characterize the decision-making process. The drift rate, v, captures the speed at which evidence toward a decision is accumulated; that is, drift rates are typically faster in easy tasks (e.g., congruent trials) than in difficult tasks (e.g., incongruent trials). The threshold, a, represents the degree of response caution, with higher values indicating that decision-makers adopt a more stringent standard of evidence before making a decision (and lower values indicating that less evidence suffices). Non-decision time, t, captures the amount of time employed in parsing the relevant stimuli before initiating the decision process, plus executing the required motor response (e.g., pressing a button in an experiment). Finally, the bias parameter, z, captures whether people are initially biased toward either response option, before having accumulated any evidence toward a decision.

In short, the goal of this work is to investigate whether tension between the text and purpose of written rules generates interference effects similar to those found on basic conflict tasks. We predict that interference will manifest in longer reaction times when applying rules to hard or 'incongruent' cases (i.e., in which someone's behavior violates the rule's text yet complies with its purpose, or *vice versa*) than in easy or 'congruent' cases (i.e., in which a behavior violates both the rule's text *and* its purpose, or complies with both). Derivatively, our studies aim to understand the cognitive processes that underlie people's application of rules through the lens of drift diffusion modeling.

Experiment 1: Violating and complying with rules

Experiments 1a and 1b were designed as a first exploration of judgment and reaction time data using the drift diffusion framework. The experiments are identical except for the test question ("did this person violate the rule?" vs. "did this person comply with the rule?"). Both experiments were preregistered (Exp. 1a: https://aspredicted.org/mm6ny.pdf, Exp. 1b: https://aspredicted.org/4jy82.pdf). We implemented the tasks in *jsPsych* (https://www.jspsych.org/7.3/), and hosted them on cognition.run. Data were analyzed using R and RStudio, as well as the *hddm* package in python (for hierarchical drift diffusion modeling). Materials, data, and code for all experiments presented in this paper are available at https://osf.io/wsejx/. Online demos of all experiments are available at https://neeleengelmann.github.io/2024_cogsci_rules/.

Experiment 1a: Violating rules

Design, Material and Procedure We used a 2 (text violation: yes vs. no) × 2 (purpose violation: yes vs. no) × 3 (items) × 8 (scenario) design, with all manipulations administered within subjects. Thus, each subject saw 12 trials in each of the 8 scenario blocks, for a total of 96 trials. The order of scenario blocks and the order of trials within blocks was randomized. Each block began with the presentation of a rule and its purpose, e.g.: "To keep her house clean, Mary announced: No one may wear shoes in the house." On the following pages, participants were presented with example behaviors (one per page) that either violated the rule's text e.g., "A guest wears his new sneakers and the floor stays clean"), its purpose (e.g., "A barefoot guest has a bleeding toe and stains the carpets"), both (e.g., "A guest wears muddy sneakers and dirties the carpets"), or neither (e.g., "A guest walks around barefoot and keeps the carpet clean"). On each page, we asked "Did this person violate the rule?", and participants were instructed to answer yes or no as fast as possible by pressing either the e key or the i key on their keyboard. Assignment of response keys was counterbalanced across participants. After a time limit of 8 seconds without a response, a trial was recorded as invalid. A fixation cross was presented between trials with a random duration between 250 and 2,000 ms. Before the main experiment, participants completed a practice block (without time pressure or incongruent trials) to familiarize themselves with the layout and task. Participants had the opportunity to take a break after each block. The experiment ended with the assessment of demographic variables and a debriefing.

Participants 119 participants completed the survey on *prolific.com* (M_{age} = 35.9, SD_{age} = 14 years, 55% women, 44% men, 1% non-binary or no answer). Inclusion criteria (for all three experiments) were: being a native English speaker, not having participated in previous studies using similar materials, and an approval rate of at least 90% on previous tasks on the platform. Participants received a compensation of £1.50 for an estimated ten minutes of their time.

Results and Discussion See Figure 1a for an overview. As expected, participants overwhelmingly indicated that the rule was violated when the target behavior conflicted with both its text and its purpose (95% yes, 95% CI: 94-96%), and that the rule was *not* violated (4% yes, 95% CI: 3-5%) in absence of either text or purpose violation. Replicating previous results (Struchiner et al., 2020; Hannikainen et al., 2022), violation of text alone more often led people to judge that the rule was violated (65% yes, 95% CI: 63-67%) than violation of purpose alone (26% yes, 95% CI: 23-29%). The data were best described by a mixed logistic model with fixed effects of text ($\chi^2_{df=1}$ = 6700.3, p<.001) and purpose ($\chi^2_{df=1}$ = 1646.7, p<.001), in addition to random intercepts for participant and scenario.

Turning to response times, participants reacted faster to congruent cases of violation (median = 2,307 *ms*) or compliance (median = 2,285 *ms*) than to conflict cases (text violation: median = 2,999 *ms*, purpose violation: median = 2,658 *ms*). Accordingly, the data were best described by a

linear mixed model containing fixed effects for text ($\chi^2_{df=1}$ = 96.39, $p<.001$), purpose ($\chi^2_{df=1}$ = 40.97, $p<.001$), and their two-way interaction ($\chi^2_{df=1}$ = 740.75, $p<.001$) in addition to random intercepts for participant and scenario.[2]

Visual inspection of Figure 1a revealed that 'yes' responses tended to be faster than 'no' responses (full violation: Δ_{RT} = 1,063 *ms*, text-only violation: Δ_{RT} = 473 *ms*, purpose-only violation: Δ_{RT} = 124 *ms*), except in the compliance condition (Δ_{RT} = -196 *ms*). As an exploratory analysis, we thus tested whether adding a three-way interaction between text, purpose, and response (yes vs. no) improved model fit, and found that it did ($\chi^2_{df=4}$ = 158.56, $p<.001$).

Next, we ran a series of drift diffusion models in the *hddm* Python library (Wiecki, Sofer, & Frank, 2013), with 10,000 samples, a burn-in rate of 10% and a thinning factor of 3. We forced a constant boundary separation across case-types, under the plausible assumption that participants would not adjust their degree of response caution flexibly on each trial, given that trial type was randomized and thus unpredictable. On the basis of the Deviance Information Criterion, and visual evidence of convergence in the trace plots (see Online Materials), we compared models that allowed the drift rate, non-decision time, and/or bias to vary across case-types.

The best-fitting model revealed a positive bias (z = .53, 95% HDI [.53, .54]), i.e., a bias toward the 'Yes' response, and a boundary separation of a = 3.16 (95% HDI [3.07, 3.25]). Non-decision times differed modestly across conditions, between 1,014 and 1,171 *ms*, as shown in Figure 2C. Meanwhile, drift rates were faster in congruent cases than incongruent cases (see Figure 2D)—both in the comparison between full violation (v = 1.02, 95% HDI [0.95, 1.10]) and literal violation (v = 0.24, 95% HDI [0.17, 0.30]), and between full compliance (v = -1.18, 95% HDI [-1.25, -1.11]) and literal compliance (v = -0.51, 95% HDI [-0.58, -0.45]). Relaxing the assumption of constant boundary separation did not qualitatively affect this pattern of results (see Online Appendix 1).

Experiment 1b: Complying with rules

In Experiment 1b, we sought to generalize the behavioral results of Experiment 1a to judgments of *compliance*, and to assess the degree of convergence between the drift diffusion parameters across both experiments. Additionally, Experiment 1b allowed us to disambiguate whether the faster 'yes' responses in Experiment 1a reflect general acquiescence (a preference for affirmative responses), or instead an "accusatory stance" (preferring to convict than acquit, independently of how judgments are elicited).

Design, Material and Procedure This experiment was identical to Experiment 1a in every respect, except for the test question. We now asked: 'Did this person comply with the rule?´.

[2]Deviating from our preregistration, we report a *linear* model of reaction times due to lack of convergence in an ordinal model.

Participants 124 participants completed the survey on *prolific.com* (M_{age} = 36, SD_{age} = 12.3 years, 49% women, 48% men, 3% non-binary or no answer). Participants received a compensation of £1.50 for an estimated ten minutes of their time.

Results and Discussion See Figure 1A for an overview. With the inverse dependent measure, participants almost always indicated that agents in congruent cases complied with the rule when neither text nor purpose had been violated (93% yes, 95% CI: 92-94%), and that they did not comply with the rule when violating both text and purpose (6% yes, 95% CI: 5-7%). Mirroring the pattern of results in Experiment 1a, violations of purpose alone were more often judged in compliance with the rule (73% yes, 95% CI: 70-76%) than violations of text (28% yes, 95% CI: 26-30%). The data were best described by a mixed logistic model with fixed effects for text ($\chi^2_{df=1}$ = 45.16, $p<.001$), purpose ($\chi^2_{df=1}$ = 26.75, $p<.001$), and their two-way interaction ($\chi^2_{df=2}$ = 776.23, $p<.001$), in addition to random intercepts for scenario and participant.

As in Experiment 1a, participants were faster to decide congruent cases of rule violation (median = 2,335 *ms*) and compliance (median = 2,296 *ms*) than incongruent cases (purpose violation: median = 2,710 *ms*, text violation: median = 2,940 *ms*). Across all conditions, "no" responses were now faster than "yes" responses, though the size of this median difference varied across conditions (violation: Δ_{RT} = 346 *ms*, compliance: Δ_{RT} = 158 *ms*, text violation: Δ_{RT} = 534 *ms*, purpose violation: Δ_{RT} = 201 *ms*). The data were best described by a linear mixed model with fixed effects for text ($\chi^2_{df=1}$ = 45.16, $p<.001$), purpose ($\chi^2_{df=1}$ = 26.75, $p<.001$), response (yes vs. no) ($\chi^2_{df=1}$ = 75.53, $p<.001$), as well as the text × purpose interaction ($\chi^2_{df=1}$ = 700.7, $p<.001$), and the three-way text × purpose × response interaction ($\chi^2_{df=3}$ = 41.31, $p<.001$), in addition to random intercepts for participant and scenario. Thus, the faster violation judgments observed in Experiment 1a were not due to acquiescence, but rather to a greater ease in indicating that a target behavior is against the rule (rather than in compliance).

Again, we fit a series of hierarchical drift diffusion models, with a constant boundary separation across case-types, and applied the same criteria for model specification and selection as in Experiment 1a. The best-fitting model revealed a bias toward the 'no' response (z = .46, 95% HDI [.45, .47]) with a boundary separation of a = 3.23 (95% HDI [3.11, 3.36]). Non-decision times again differed modestly across conditions (see Figure 2C), ranging between 1019 and 1120 *ms*. Drift rates were substantially slower for literal violation (v = -0.36, 95% HDI [-0.44, -0.29]), than full violation cases (v = -1.03, 95% HDI [-1.12, -0.95]), and for literal compliance (v = 0.55, 95% HDI [0.46, 0.62] than full compliance cases (v = 1.15, 95% HDI [1.05, 1.22]). Relaxing the assumption of constant boundary separation did not qualitatively affect this pattern of results (see Online Appendix 2).

Thus, Experiments 1a and 1b revealed an interference ef-

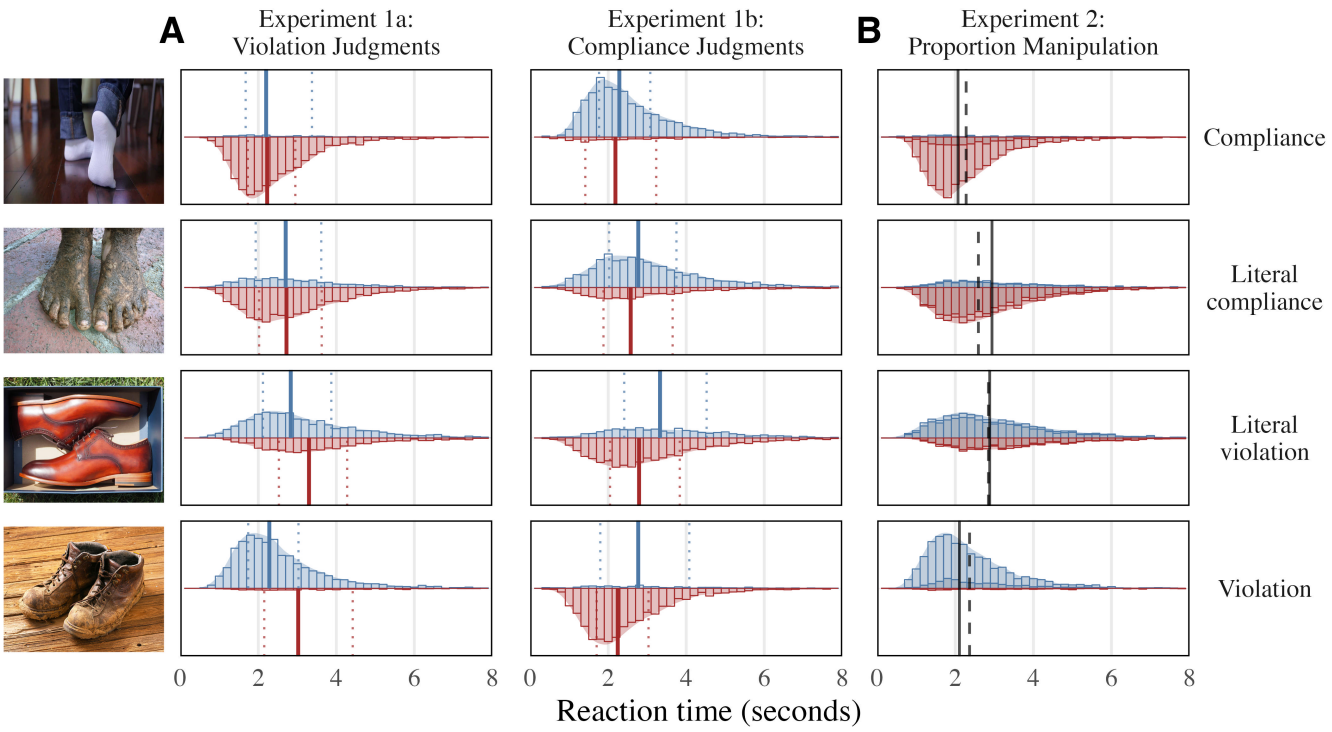

Figure 1: **Reaction Time Distributions in Experiments 1 and 2.** The figure displays grouped histograms of 'Yes' (**blue**) and 'No'(**red**) responses, separately for each experiment and case type. Overlaid vertical lines represent (A) the median, 1st and 3rd quartiles, and (B) the median in the congruence dominant (solid) and incongruence dominant (dashed) conditions.

fect on reaction times. In both experiments, drift diffusion modeling uncovered that this effect was primarily attributable to changes in the average drift rate across case-types. Non-response times, by contrast, independently contributed to longer reaction times only in cases of literal violation. In light of previous validation studies on the drift diffusion model, this pattern of results can be interpreted as revealing cognitive conflict in the application of rules to hard cases. Additionally, both experiments uncovered shorter reaction times for judgments of conviction than of acquittal and, accordingly, the drift diffusion model retrieved a bias toward (i.e., greater initial proximity to) the decision boundary representing conviction in both experiments.

Experiment 2: Varying the proportion of incongruent trials

In Experiment 1, case-types were presented in a random order and with an equal probability, precluding the possibility that participants would adapt their response strategy to characteristics of the task. Accordingly, we assumed that the boundary separation parameter—a relative stable aspect of participants' response strategy—would remain constant across trials.

The purpose of Experiment 2 is to further validate the application of the drift diffusion framework to rule-based decision-making by leveraging a manipulation of the proportion of congruency trials—known, in the cognitive control literature, to influence the magnitude of the interference effect. Typically, a low proportion of incongruent trials increases the interference effect (producing longer reaction times and more errors on incongruent trials), while a high proportion of incongruent trials reduces them (Bugg & Crump, 2012). Drift diffusion modeling has shown, in this context, that block-wide manipulations of the congruency proportion allow participants to adapt their degree of response caution to changes in the likelihood of encountering congruent vs. incongruent trials across experimental blocks. Our goal was to test whether and how boundary separation is affected by this manipulation, and to study the behavior of the remaining parameters under these conditions. The experiment is preregistered at https://aspredicted.org/zi8qe.pdf.

Design, Material and Procedure The experiment was identical to the previous studies in all aspects except we introduced a manipulation of the proportion of congruent (i.e., violation and compliance) vs. incongruent (i.e., literal violation and literal compliance) trials per rule. Participants now saw two blocks consisting of four rules each: a block with a high proportion of congruent trials (10 out of 12 trials for each rule; 5 violation, 5 compliance, 1 literal violation, and 1 literal compliance) and another block with a low proportion of congruent trials (2 out of 12 trials for each rule). The or-

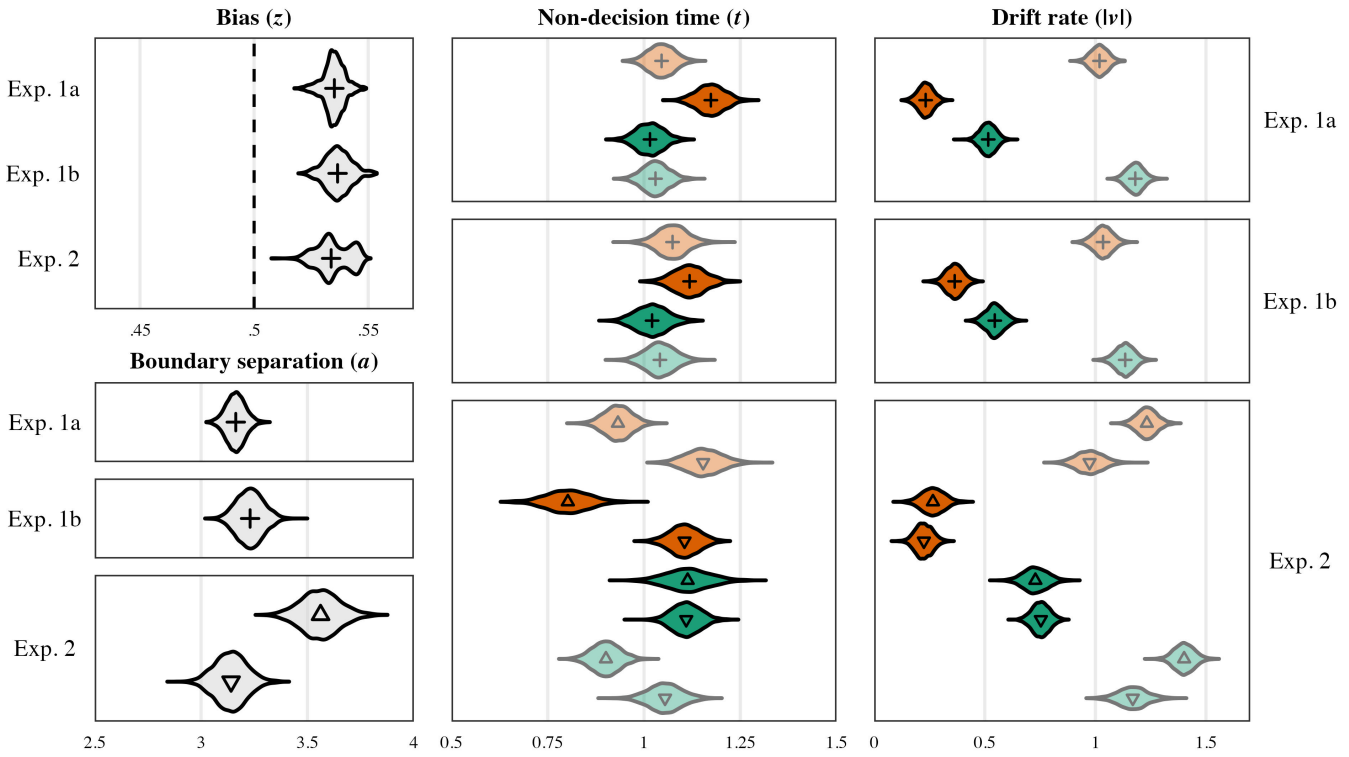

Figure 2: **Posterior Distributions of the HDDM Parameter Estimates in Experiments 1 and 2.** (A) Bias and (B) boundary separation in each experiment, with boundary separation displayed separately for congruent dominant (△) and incongruent dominant (▽) blocks. (C) Non-decision times and (D) absolute drift rates are displayed separately for congruent (**light shade**) and incongruent (**dark shade**) cases of violation (**brown**) and compliance (**green**). *Note*: bias in Experiment 1b has been reversed to facilitate visual comparison.

der of the congruency proportion blocks was randomized between subjects. To ensure that no rule appeared twice for any given participant, the rules that made up the first and second blocks were randomly drawn from the set of eight rules without replacement. Participants were given no indication of the manipulation of congruency proportion, and all instructions remained identical to the previous studies. The test question was "Did this person violate the rule?", as in Experiment 1a.

Participants 121 participants completed the survey on *prolific.com* (M_{age} = 35.2, SD_{age} = 13.3 years, 51% women, 48% men, 1% non-binary or no answer) and received £1.50 as compensation for an estimated ten minutes of their time.

Results and Discussion Replicating the pattern of results observed in previous experiments (see Figure 1b), literal violations were seen as rule violations less often (65% yes, 95% CI: 63-67%) than full violations (95% yes, 95% CI: 94-96%), while cases of literal compliance were seen as violating the rule more often (18% yes, 95% CI: 15-21%) than cases of full compliance (4% yes, 95% CI: 3-5%). The data were best described by a model that contained fixed effects of text ($\chi^2_{df=1}$ = 6857.6, p<.001), purpose ($\chi^2_{df=1}$ = 1183.2, p<.001), proportion of congruency ($\chi^2_{df=1}$ = 9.0, p = .003), and the text × purpose × proportion of congruency interaction ($\chi^2_{df=3}$ = 12.11, p = .007), in addition to random intercepts for participant and scenario. [3]

For our preregistered analysis of reaction times, we collapse the types of trials into the class of congruent (full violation and compliance) and incongruent (text-only violation and purpose-only violation) trials. Responses to congruent trials were faster than responses to incongruent trials in both congruency proportion blocks, but the median difference was larger when incongruent trials were rare (Δ_{RT} = 822 *ms*) than when they were frequent (Δ_{RT} = 395 *ms*). Accordingly, the data were best described by a linear model containing fixed effects of congruency ($\chi^2_{df=1}$ = 476.37, p<.001), congruency proportion ($\chi^2_{df=1}$ = 10.57, p = .001), and the congruency × proportion interaction ($\chi^2_{df=1}$ = 63.06, p<.001), in addition to random intercepts for participant and scenario.

In our drift diffusion models, we now freed the boundary

[3] The interaction effect involving the congruency proportion manipulation reflected a weak tendency toward elevated rule violation judgments in congruent-dominant blocks for all case-types (*OR*s ranging from 1.06 to 1.38) except full compliance (where *OR* = 0.52). However, given the magnitude of these effects, we take them to be theoretically irrelevant.

separation parameter, allowing congruency proportion to impact individuals' response caution (as well as non-decision times and drift rates). Boundary separation was larger in blocks with a majority of congruent trials (a = 3.56 (95% HDI [3.11, 3.36]) than in blocks with a majority of incongruent trials (a = 3.14 (95% HDI [3.00, 3.29]). The proportion manipulation also influenced non-decision times, which were generally longer in the incongruent dominant block (between 1,058 and 1,146 *ms*) than in the congruent dominant block (between 792 and 922 *ms*), except in literal violation cases where the difference was negligible (from 1,111 to 1,114 *ms*; see Figure 2d).

The proportion manipulation revealed differential effects on evidence accumulation during congruent versus incongruent cases. Specifically, drift rates during congruent trials were faster in blocks where they were frequent (violation: v = 1.24, 95% HDI [1.16, 1.33]; compliance: v = -1.40, 95% HDI [-1.48, -1.30]) than in blocks where they were infrequent (violation: v = 0.99, 95% HDI [0.86, 1.11]; compliance: v = -1.16, 95% HDI [-1.30, -1.04]). By contrast, the proportion of congruency had no effect on drift rates for either literal violation (*high congruency* v = 0.23, 95% HDI [0.16, 0.31]; vs. *low congruency* v = 0.27, 95% HDI [0.18, 0.37]) or literal compliance (*high congruency* v = -0.75, 95% HDI [-0.83, -0.67]; *low congruency* v = -0.72, 95% HDI [-0.84, -0.61]) cases. Lastly, the best-fitting model revealed a prosecution bias (z = .53, 95% HDI [.52, .54]) as in Experiment 1.

Thus, we reproduced the widely-documented *proportion congruency effect* (Bugg & Crump, 2012) on reaction times in the domain of rule violation judgments: namely, the pattern of interference was *amplified* during blocks in which most cases were congruent. A drift diffusion model revealed that these high congruency blocks fostered a greater emphasis on accuracy than did low congruency blocks (in which most cases were incongruent), while also accelerating evidence accumulation *selectively* on congruent trials. Meanwhile, the manipulation of congruency proportion did not affect drift rates on incongruent trials.

General Discussion

Multiple recent studies provide evidence that moral cognition can come into conflict with people's literal understanding of legal and communal rules (Struchiner et al., 2020; Turri, 2019; Hannikainen et al., 2022). Meanwhile, developments in drift diffusion modeling have afforded a clearer understanding of the cognitive processes that facilitate decision-making on basic perceptual tasks evoking response competition. In this work, we brought together these parallel lines of research, and contributed to ongoing efforts to apply drift diffusion modeling to higher-order reasoning tasks (Cohen & Ahn, 2016; Yu, Siegel, Clithero, & Crockett, 2021; Siegel, van der Plas, Heise, Clithero, & Crockett, 2022; Engelmann & Hannikainen, 2024).

Behaviorally, we replicated typical effects of interference and congruency proportion observed in basic conflict tasks (e.g., Stroop and Flanker tasks) (Bugg & Crump, 2012). Cases in which the application of purposivist and textualist standards would yield opposing verdicts produced longer reaction times, and this tendency was amplified when the proportion of incongruent trials was low.

Drift diffusion models indicated that the interference effect in both Experiments 1a and 1b was attributable to a slower rate of evidence accumulation (or drift rate), and this reduction persisted when experimentally elevating the proportion of incongruent trials in Experiment 2. Thus, greater exposure to incongruent cases did not appear to facilitate evidence accumulation on incongruent trials—which further supports the interpretation of slower drift rates as a marker of response competition. Additionally, the bias parameter indicated that the starting point of evidence accumulation was closer to the conviction boundary than the acquittal boundary—a tendency that dovetails with various asymmetries in the attribution of third-party blame versus praise (Guglielmo & Malle, 2019), and explains the tendency for conviction to occur faster than acquittal in all three experiments.

In sum, our present studies exploited a fruitful, though imperfect, analogy between letter vs. spirit conflicts in statutory interpretation and lower-level forms of response competition. Still, whether domain-general mechanisms of conflict monitoring and control guide decision-making across these disparate tasks cannot be gleaned from the present evidence alone. Additionally, understanding the trajectory of evidence accumulation over time in a *time-varying* framework (Ulrich, Schröter, Leuthold, & Birngruber, 2015) may provide novel insights into the onset of competing textualist and purposivist responses (see also Flanagan et al. (2023)). Nevertheless, our studies raise the prospect of better understanding how people reason about everyday transgressions of written rules and laws by applying insights from the the drift diffusion framework.

References

Almeida, G., Struchiner, N., & Hannikainen, I. R. (2023). Rule is a dual character concept. *Cognition*, *230*, 105259.

Awad, E., Levine, S., Loreggia, A., Mattei, N., Rahwan, I., Rossi, F., ... Kleiman-Weiner, M. (2022). When is it acceptable to break the rules? Knowledge representation of moral judgement based on empirical data. Retrieved from https://arxiv.org/abs/2201.07763

Bridgers, S. E. C., Taliaferro, M., Parece, K., Schulz, L., & Ullman, T. (2023). Loopholes: A window into value alignment and the communication of meaning. Retrieved from https://osf.io/preprints/psyarxiv/cnxzv

Bugg, J. M., & Crump, M. J. (2012). In support of a distinction between voluntary and stimulus-driven control: A review of the literature on proportion congruent effects. *Frontiers in Psychology*, *3*, 367.

Cohen, D. J., & Ahn, M. (2016). A subjective utilitarian theory of moral judgment. *Journal of Experimental Psychology: General*, *145*(10), 1359.

Engelmann, N. (2023). Murderer at the door! To lie or to mislead? In A. Wiegmann (Ed.), *Lying, fake news, and bullshit.* Bloomsbury. (In press)

Engelmann, N., & Hannikainen, I. R. (2024). Does moral valence influence the construal of alternative possibilities? *Possibility Studies & Society*. (in press)

Flanagan, B., Almeida, G., Struchiner, N., & Hannikainen, I. R. (2023). Moral appraisals guide intuitive legal determinations. *Law and Human Behavior*, *47*(2), 367.

Fuller, L. L. (1957). Positivism and fidelity to law – A reply to Professor Hart. *Harvard Law Review*, *71*, 630.

Guglielmo, S., & Malle, B. F. (2019). Asymmetric morality: Blame is more differentiated and more extreme than praise. *PloS one*, *14*(3), e0213544.

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

Kwon, J., Tenenbaum, J., & Levine, S. (2022). Flexibility in moral cognition: When is it okay to break the rules? In J. Culbertson, H. Rabagliati, V. Ramenzoni, & A. Perfors (Eds.), *Proceedings of the annual meeting of the cognitive science society* (Vol. 44, pp. 905–911).

Kwon, J., Zhi-Xuan, T., Tenenbaum, J., & Levine, S. (2023). When it's not out of line to get out of line: The role of universalization and outcome-based reasoning in rule-breaking judgments. Retrieved from https://osf.io/preprints/psyarxiv/n8bjr

Myers, C. E., Interian, A., & Moustafa, A. A. (2022). A practical introduction to using the drift diffusion model of decision-making in cognitive psychology, neuroscience, and health sciences. *Frontiers in Psychology*, *13*, 1039172.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873–922.

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review*, *111*(2), 333.

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences*, *20*(4), 260–281.

Schauer, F. (1991). *Playing by the rules: A philosophical examination of rule-based decision-making in law and in life*. Clarendon Press.

Schelling, T. C. (1960). *The strategy of conflict.* Harvard University Press.

Siegel, J. Z., van der Plas, E., Heise, F., Clithero, J. A., & Crockett, M. (2022). A computational account of how individuals resolve the dilemma of dirty money. *Scientific Reports*, *12*(1), 18638.

Struchiner, N., Hannikainen, I. R., & de Almeida, G. d. F. (2020). An experimental guide to vehicles in the park. *Judgment and Decision Making*, *15*(3), 312–329.

Sunstein, C. R. (2005). Moral heuristics. *Behavioral and Brain Sciences*, *28*(4), 531–541.

Turri, J. (2019). Excuse validation: A cross-cultural study. *Cognitive Science*, *43*(8), e12748.

Ulrich, R., Schröter, H., Leuthold, H., & Birngruber, T. (2015). Automatic and controlled stimulus processing in conflict tasks: Superimposed diffusion processes and delta functions. *Cognitive Psychology*, *78*, 148–174.

Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: An empirical validation. *Memory & Cognition*, *32*, 1206–1220.

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). Hddm: Hierarchical bayesian estimation of the drift-diffusion model in python. *Frontiers in Neuroinformatics*, 14.

Yu, H., Siegel, J. Z., Clithero, J. A., & Crockett, M. J. (2021). How peer influence shapes value computation in moral decision-making. *Cognition*, *211*, 104641.