Coordination Favors Legal Textualism by Suppressing Moral Valuation

Ivar R. Hannikainen, Kevin P. Tobia, Guilherme da F. C. F. de Almeida, Noel Struchiner, Markus Kneer, Piotr Bystranowski, Vilius Dranseika, Niek Strohmaier, Sammy Bensinger, Kristina Dolinina, Bartosz Janik, Eglė Lauraitytė, Michael Laakasuo, Alice Liefgreen, Ivars Neiders, Maciej Próchnicki, Alejandro Rosas, Jukka Sundvall, & Tomasz Żuradzki

* Correspondence concerning this article should be addressed to: Ivar Rodríguez Hannikainen, Department of Philosophy I, Faculty of Psychology, Cartuja Campus, Universidad de Granada (Spain), 18011. Contact: ivar@ugr.es

Word count (excluding References and Appendices): 3840

Abstract

A cross-cultural survey experiment revealed a general tendency to rely on a rule's text over its purpose when deciding which acts violate the rule. This tendency's strength varied markedly across (k = 13) field sites, owing to cultural differences in the impact of moral appraisals on judgments of rule violation. Next, we observed that legal experts were more strongly inclined to disregard their moral evaluations of the acts altogether, and they consequently demonstrated stronger textualist tendencies than did laypeople. Finally, we examined a plausible mechanism for the emergence of textualism in a two-player coordination game: Incentives to coordinate without communicating reinforced participants' reliance on rules' literal meaning. Together, these studies (total N = 5109) help clarify the origins and allure of legal textualism. While diverse legal actors may have varied personal assessments of rules' moral purposes, rules' literal meanings serve as clear *focal points*—easily identifiable points of agreement that enable coordination among diverse agents and judges.

Coordination Favors Legal Textualism by Suppressing Moral Valuation

Legal theorists have long interrogated the relationship between legal rules' literal meaning and purpose. In many cases, a rule's literal meaning and purpose support the same rule application. For example, for the purpose of reducing intoxicated driving, all fifty states have passed "zero tolerance" alcohol consumption laws for young drivers. If a person below age twenty-one drives with detectable alcohol in their bloodstream, their license is suspended (Fell et al. 2016). Both the literal meaning and purpose of this rule support suspending the license of a nineteen year-old who drinks three martinis before driving.

But sometimes a rule's literal meaning and purpose support opposing rule applications (Schauer, 1991). Rules are imperfectly *specific*: a driver under the influence of ecstasy could pose an even larger threat to road safety. Call this an "underinclusion case"; the rule's literal meaning fails to capture an act that violates the rule's purpose. They are also imperfectly *sensitive*: rinsing with an alcohol-based mouthwash might result in a positive test result, without heightening the risk of accidents. Call this an "overinclusion case"; the rule's literal meaning captures an act that is not actually a violation of the rule's purpose.

When evaluating these two hypothetical cases, it is abundantly clear whose behavior is *morally* worse: We condemn the first agent's reckless conduct and exonerate the second. This distinction arises early in development (Nucci & Turiel, 1978), as children abandon the uncritical submission to authority and autonomously reason about deeper ethical principles (Kohlberg, 1981; Sripada & Stich, 2005), and plausibly reflects *model-based* reasoning (Cushman, 2013; Miller, Hannikainen & Cushman, 2014), over the probability and magnitude of negative outcomes—most notably, harm to third parties.

Now consider a different question. Which of these two acts violates the relevant statute? Is it the first, which jeopardizes the zero tolerance policy's *purpose* of saving lives (Barak, 2005; Fuller, 1958); or is it the second, which conflicts with the *literal* meaning of the rule (Hart, 1958; Schauer, 1988)? Existing empirical work shows that, when evaluating rules and statutes, laypeople consider the underlying purpose (whether the driver posed a threat to road safety) but pay closer attention to the rule's literal meaning (i.e., whether the driver ingested alcohol; see Bregant, Welbery & Shaw, 2019; Garcia, Chen, & Gordon, 2014; Struchiner, Hannikainen, & Almeida, 2020; Turri & Blouw, 2015). As a result, most laypeople conclude that *overinclusion* cases (i.e., driving after rinsing with alcohol-based mouthwash) violate the rule while *underinclusion* cases (i.e., driving under the influence of ecstasy) do not.

This form of interpretation—relying on a rule's text over its purpose—reflects a "lay textualism." In the United States, textualism is one of the leading theories of legal interpretation, particularly at the Supreme Court (Krishnakumar 2022), and it has substantial acceptance in the legal academy as a form of statutory interpretation (Martinez & Tobia 2022). In this article, we refer to the reliance on a rule's literal meaning over its apparent purpose as "textualist" interpretation. But some sophisticated forms of textualism advise looking beyond a rule's literal meaning, taking into account context and pragmatics in interpretation of the text (Manning 2005). Nevertheless, the contrast between literal meaning and purpose is a longstanding one and captures an important debate between some forms of textualism and purposivism. As a leading textualist puts it, "texts should be taken at face value—with no implied extensions of specific texts or exceptions to general ones—even if the legislation will then have an awkward relationship to the apparent background intention or purpose that produced it." (Manning 2005: 428).

In the present work, we evaluation rule violation judgments in a large-scale cross-cultural experiment. Despite remarkable cultural variability, we document a dominant tendency toward textualist interpretation across thirteen countries. What might lead people to disregard their moral appraisals and adhere to the literal scope of a rule when assessing the legality of particular acts?

One possibility is that a focus on the rules' text constitutes a *heuristic* (Sunstein, 2005). In naturalistic contexts, a model-based evaluation of the target act can be rife with uncertainty—for instance, if the rule's purpose is unknown or the subject of disagreement. Even a simple rule that states that "no busses may enter the park" admits of various purposes; is the purpose to promote cleanliness, improve safety, decrease noise, or reduce environmental pollutants. Comparatively, determining whether the target act violates a rule's text is often cognitively simpler and less error prone (Mousavi & Gigerenzer, 2017), enabling individuals to assess a target act's value without modeling and evaluating its ensuing outcomes. However, previous experimental evidence speaks against this purely heuristic role: Rendering the rules' purpose explicit and the target acts' outcomes easily evaluable does not eliminate people's tendency toward textualist interpretation (Struchiner et al., 2020).

Here we pursue a distinct explanation rooted in the notion of a *focal point* (Schelling, 1960). Research in game theory has shown that, in cooperative contexts, people can reach an equilibrium solution even in the absence of communication, *if* they align their choices with some arbitrary but salient element of the context *and* as long as each party knows that the other is trying to do the same (e.g., when two bikers swerve right to avoid a head-on collision).

Let us suppose, then, that—in legal reasoning contexts—agents want to cooperate in determining the scope of a given rule. For instance, judges may want to consider how citizens

understand legal rules to promote 'fair notice'. Equally, citizens may wish to discern how authorities understand the scope of a rule—i.e., what does and does not constitute a transgression—in order to avoid punishment. Our final experiment provided evidence that a rule's literal meaning constitutes a focal point for interpretation (see Figure 1).

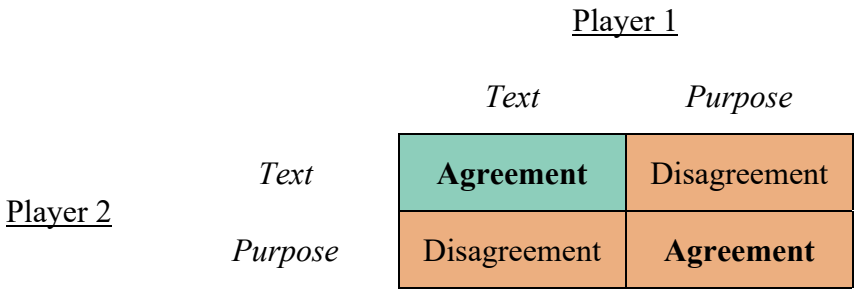

Figure 1. Legal Interpretation as Coordination.

When participants were incentivized to align their interpretations in a coordination game, they displayed stronger textualist tendencies than when asked to freely judge the same cases. Legal training also strengthened textualist tendencies—perhaps as a result of the additional coordination incentives to which legal professionals are subject: the incentives, e.g., to consider how their peers would decide an identical case in order to issue consistent verdicts and preserve the rule of law, or to avoid issuing rulings that may be criticized or reversed by judges in higher courts.

Results

Our first sequence of analyses examined the responses of laypeople (i.e., who reported no legal training). Manipulation checks confirmed that overinclusive cases were seen as violating the literal meaning more than were underinclusive cases, $B = 2.25$, $t_{(1896)} = 23.80$, while underinclusive cases were seen as more morally blameworthy than overinclusive cases, $B = -3.16$, $t_{(1890)} = -38.64$, both ps $<.001$.

Turning to our main analysis, a mixed effects model of rule violation judgments revealed an effect of case-type, which was qualified by the two-way interaction with evaluation mode (see Table 1). Replicating previous evidence (Struchiner et al., 2020), overinclusive cases ($M = 4.26$) were more likely to be considered rule violations than were underinclusive cases ($M = 3.77$), $B = 0.49$, $t = 6.62$, $p < .001$ (see also Figure 2A).

Table 1. *Mixed Effects Models: Results.*

		Laypeople				Lawyers			
		F	*df*s	*p*	η_p^2	*F*	*df*s	*p*	η_p^2
Pre-Registered Model	Type	44.36	(1, 3692)	< .001	.012	97.84	(1, 766)	< .001	.113
	Evaluation Mode	0.72	(1, 3689)	.54	.000	4.59	(1, 766)	.032	.006
	Type×Mode	10.59	(1, 3691)	.001	.003	2.61	(1, 767)	.106	.003
Exploratory Model	Country	5.08	(12, 3674)	< .001	.017	4.02	(3, 763)	.007	.016
	Type×Country	10.85	(12, 3674)	< .001	.036	7.57	(3, 763)	< .001	.029

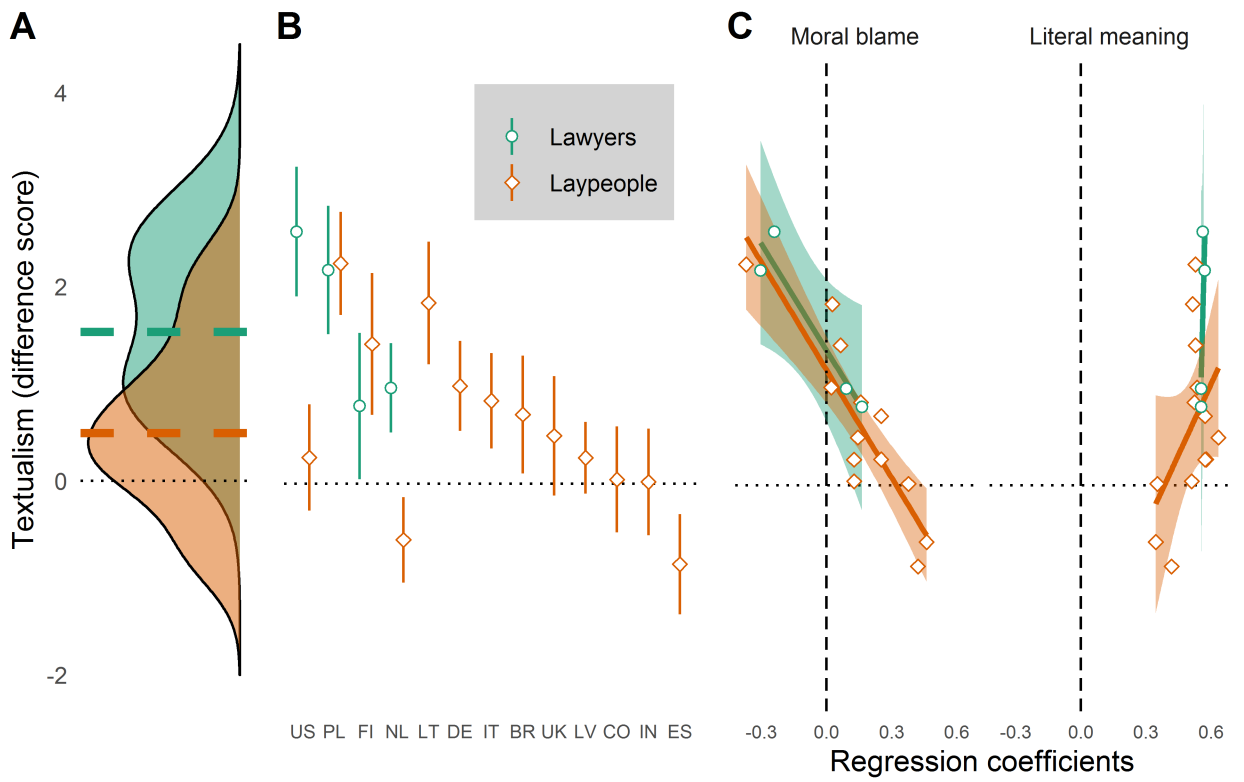

Figure 2. Textualism Effects among Laypeople and Lawyers: (A) Grouped density plot. (B) National means and 95% confidence intervals. (C) National means against multiple regression coefficients of moral blame and literal meaning.

Regressing rule violation judgments on literal meaning and moral blame demonstrated that both predictors were significant: In standardized terms, the effect of literal meaning judgments, $\beta = 0.54$, 95% CI [0.50, 0.58], $t = 25.31$, was over three times stronger than the effect of moral condemnation, $\beta = 0.15$, 95% CI [0.11, 0.19], $t = 7.63$, both $ps < .001$. In line with previous evidence (Struchiner et al., 2020; Garcia et al., 2014), laypeople across cultures demonstrated a hybrid approach to interpretation which reflected both moral and literal meaning criteria when applying rules to cases.

Understanding Cultural Variation

To examine the degree of cultural variation, next we treated country as a fixed effect in the primary model. This analysis revealed substantial variation in rule violation judgments across countries, as indicated by the country×case-type interaction, $F_{(11, 3474)} = 11.78$, $p < .001$, $\eta_p^2 = .036$. The simple effect of case-type revealed that a tendency toward textualism emerged in Brazil ($p = .022$), Finland, Germany, Italy, Lithuania, and Poland (ps <.001), but not in Colombia ($p = .88$), India ($p = .95$), Latvia ($p = .16$), the United Kingdom ($p = .12$) or the United States ($p = .34$). Furthermore, the effect reversed in two countries, Spain ($p = .002$) and the Netherlands ($p = .010$). Figure 2B displays the magnitude of this simple effect of case-type (or *textualism score*) for each national site: A positive value represents greater rule violation judgments in overinclusion cases than underinclusion cases.

Correspondingly, we also observed substantial moderation of the effect of moral blame on rule violation judgments, $F_{(12, 1832)} = 10.21$, $p < .001$, $\eta_p^2 = .066$. Meanwhile, the interaction between country and literal meaning was marginally significant, $F_{(12, 1832)} = 1.68$, $p = .066$, $\eta_p^2 = .010$; and the main effect of country was non-significant, $F_{(12, 1832)} = 0.94$, $p = .50$. In Figure 2C, we display the coefficients of moral blame and literal meaning for each country against its textualism score. The effect of moral blame varied substantially across countries, and strongly predicted textualism scores at the country level, *Spearman's* $\rho = -.83$, $p < .001$. Meanwhile, the relevance of literal meaning was not a clear predictor of cultural differences in textualism, *Spearman's* $\rho = .39$, $p = .19$. In other words, cultural differences in textualism were explained by variability in the extent to which moral evaluations (and not literal meaning) predicted rule violation judgments.

Elevated Textualism among Legal Experts

Turning to the legal expert data, manipulation checks confirmed that (i) overinclusive cases were seen as proscribed by the rule's literal meaning to a greater extent than underinclusion cases, $B = 2.39$, $t_{(377)} = 11.74$, and (ii) underinclusive cases were seen as more morally blameworthy than overinclusive cases, $B = -3.52$, $t_{(377)} = -20.29$, both ps <.001.

Our primary analysis revealed a large effect of case-type, and a small effect of evaluation mode (see statistics in Table 1). This time, the two-way interaction was not statistically significant (see Supplementary Analysis 1). The effect of case-type indicated that overinclusion cases ($M = 4.67$) were more likely to be considered violations than were underinclusion cases ($M = 3.14$), $B = 1.53$, $t = 9.91$, $p < .001$. This textualism effect emerged among legal experts in all four countries when analyzed separately: Finland ($p = .027$), the Netherlands, Poland, and the United States (remaining ps < .001).

Turning to the multiple regression analyses, the effect of literal meaning was highly significant, and in standardized terms corresponded to $\beta = 0.59$, 95% CI [0.51, 0.68], $t = 13.26$, $p < .001$. Meanwhile, contrary to laypeople, legal experts showed no effect of moral blame on rule violation judgments, $\beta = -0.06$, 95% CI [-0.13, 0.02], $t = -1.42$, $p = .16$ (see Figure 3A). As such, whereas laypeople appeared to call upon both textual and moral criteria in rule interpretation, legal experts adopted a predominantly textualist approach.

Expertise Effect: A Quasi-Experiment

In order to examine the effect of legal expertise, we recruited both legal experts and laypeople in four countries (Finland, the Netherlands, Poland, and the United States). The lay and lawyer samples revealed imbalance on every demographic measure: Lawyers were older than laypeople (Ms = 40.6 vs. 37.3), Welch's $t = 4.77$, $p < .001$, and less likely to be men (50.9%

vs. 57.4%), $\chi^2_{(df = 2)} = 9.42$, $p = .008$. In addition, the ratio of lawyers to laypeople varied across the four sites (ranging from 0.59 to 0.87 lawyers per layperson), $\chi^2_{(df = 3)} = 13.38$, $p = .004$.

In order to estimate the effect of legal expertise, we sought to eliminate extraneous, demographic differences between the lay and expert samples. So, we computed each participant's *propensity score* (Rosenbaum & Rubin, 1983; Ho et al., 2007), with higher scores indicating a greater predicted probability of being in the lawyer sample. Employing nearest-neighbor matching ($n_{pairs} = 772$) on the propensity scores, we eliminated imbalance in age (*M*s = 40.6 vs. 39.9), Welch's $t = 0.86$, $p = .39$, gender distribution (51.6% vs. 53.4%), $\chi^2_{(df = 2)} = 4.34$, $p = .11$, and country, $\chi^2_{(df = 3)} = 2.02$, $p = .57$.

In a mixed-effects model entering the expertise term, we found a two-way interaction with case-type, $F_{(1, 1532)} = 21.13$, $\eta_p^2 = .014$, $p < .001$. The simple effects of expertise revealed that lawyers were less likely to view underinclusive cases as transgressions, $B = -0.69$, $t = -4.42$, $p < .001$, and slightly more likely to judge overinclusive cases as transgressions, $B = 0.32$, $t = 2.04$, $p = .042$, than was the matched group of laypeople (see Figure 3B).

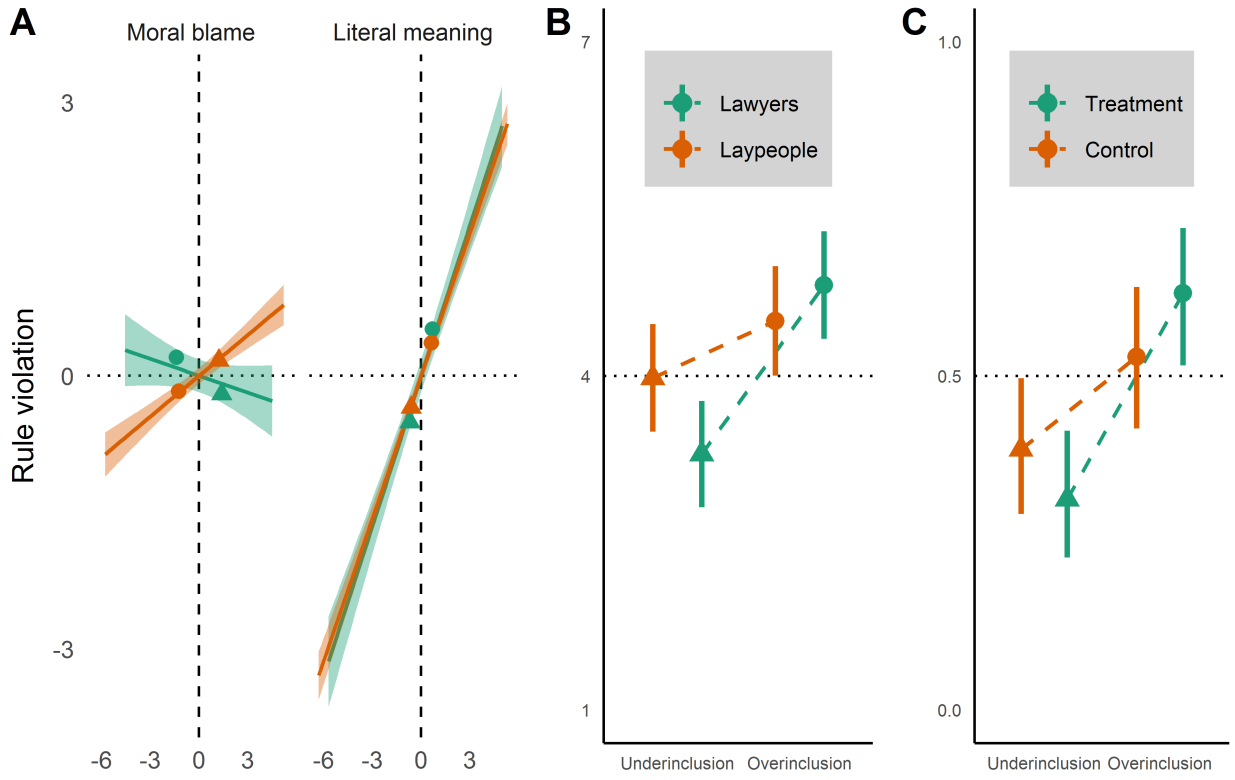

Figure 3. Rule Violation Judgments. (A) Partial residual plots by literal meaning and moral blameworthiness ratings. Mean violation judgment and 95% confidence interval for overinclusive and underinclusive cases, highlighting the effects of (B) expertise and (C) coordination incentives.

Corresponding regression analyses revealed no main effect of expertise, $F_{(1, 753)} = 0.03$, $p = .86$, or interaction with literal meaning, $F_{(1, 753)} = 2.69$, $p = .101$. An interaction with moral blame, however, did emerge, $F_{(1, 753)} = 6.83$, $p = .009$, $\eta_p^2 = .009$. Specifically, moral blameworthiness predicted violation judgments among laypeople, $B = 0.11$, 95% CI [0.02, 0.19], but not legal experts, $B = -0.05$, 95% CI [-0.13, 0.03], $t = -2.61$, $p = .009$—confirming that the experts' tendency toward greater textualism was the result of decoupling moral valuation from rule application. Supplementary Analysis 2 reveals qualitatively indistinguishable results when comparing lawyers to the entire (non-matched) pool of laypeople.

In sum, a stronger textualist response pattern was observed among lawyers than laypeople. Culture predicted variation in statutory interpretation among laypeople, by moderating the effect of moral blameworthiness (but not literal meaning) on transgression judgments. Among lawyers, cultural variability was significantly weaker, and the effect of moral blameworthiness was absent overall. Resultantly, lawyers revealed robust textualist tendencies, whereas lay samples varied in their approaches to statutory interpretation. In the following section, we pursue a potential explanation for this set of results—why textualism arises (in the face of conflicting moral attitudes), and why it might be strengthened by legal training.

Text as Focal Point in a Coordination Game

Here we present evidence that coordination motives in legal reasoning contexts may undergird people's tendency toward textualism. Our empirical prediction builds on the theoretical proposal that a rule's literal meaning serves as a *focal point* (Schelling, 1960), allowing citizens and legal officials to coordinate in the absence of communication.

To evaluate this prediction, we conducted a follow-up experiment examining people's response pattern in an incentivized, two-player coordination game. We predicted that, if a rule's literal meaning is a focal point, the incentive to coordinate interpretive judgments should further increase participants' reliance on literal meaning over purpose. We analyzed the data in a mixed-effects logistic regression with case-type, condition (Control, Game), and the case-type×condition interaction as fixed effects (treating participants and scenarios as crossed random effects). This model revealed an effect of case-type, $\chi^2_{(df = 1)} = 135.35$, and critically a case-type×condition interaction, $\chi^2_{(df = 1)} = 24.28$, both ps $< .001$. No main effect of condition was observed, $\chi^2_{(df = 1)} = 0.29$, $p = .59$.

As predicted, the interaction effect indicated that overinclusion cases (literal meaning violated, purpose not violated) were more likely to be considered transgressions in the Game condition (*prob*. = .62, 95% CI [.52, .72]) than in the Control condition (*prob*. = .53, 95% CI [.42, .63]), $OR = 1.48$, $z = 3.84$, $p < .001$. Underinclusion cases (literal meaning not violated, purpose violated) were less likely to be considered transgressions in the Game condition (prob. = .32, 95% CI [.23, .42]) than in the Control condition (prob. = .39, 95% CI [.29, .50]), $OR = 0.72$, $z = 3.15$, $p = .002$ (see Figure 3C). In sum, experimental conditions incentivizing two-player coordination strengthened laypeople's baseline textualist tendencies.

Discussion

Despite cultural variability in the impact of moral reasoning on rule interpretation, participants' general response pattern tended to reflect a more textualist approach. This tendency was greater among legal experts and in a pre-registered economic game when incentivizing participants to coordinate their interpretations without communication.

Lawyers and laypeople in our study recognized that underinclusive acts are morally blameworthy whereas overinclusive acts are not (driving after taking ecstasy is more morally blameworthy than driving after using alcoholic mouthwash). However, when reasoning about the acts' *legality*, they put forth the very opposite responses—reporting that underinclusive acts do not violate the corresponding rules, while overinclusive acts do. Lawyers' stronger textualist tendencies were partly explained by the absence of an effect of moral valuation on their rule violation judgments (Kahan et al., 2015; Tobia, 2021).

Why would people in general, and lawyers all the more, disregard their moral sense and adhere to the text of a rule when asked to adjudicate cases? The tendency to privilege a statute's literal meaning can be explained by drawing attention to rules' social dimension (Bicchieri,

2005). Some individuals devise rules, others must adhere to them (and regulate their behavior accordingly), and others yet must determine when they have been violated. If this collective practice is to be effective, the parties must share a common understanding of the rule's demands. Previous legal scholarship has theorized that the "plain meaning of a text as applied to a set of facts" can play this precise role—i.e., of serving as a coordination device (Eskridge, 1998)—and has even advocated for textualist judicial interpretation on such grounds (Leal, 2015). Our final experiment demonstrated that, when incentivized to coordinate their interpretations of rules, people without legal training strengthen their adherence to the rules' text—even though doing so incurs a moral cost.

Implications and Limitations

These results inform ongoing legal debates. In American legal interpretation, modern textualist judges increasingly aim to interpret laws in line with what those laws communicate to an *ordinary member of the public* (see, e.g., Bostock v. Clayton County, 2020). The results here suggest some support for this view: Ordinary people's understanding of legal rules is heavily informed by the rule's text.

At the same time, the coordination task results suggest that judges' real-world convergence on text might not necessarily reflect those experts' consensus about the rule's *meaning*. The same convergence could also be explained by a rational response to coordination incentives. In other words, real world coordination around rules' text might reflect experts' evaluation that text is central to meaning, or it might reflect experts' desire to coordinate around a clear focal point.

When individuals wish to coordinate their interpretations of a rule, and must consider whether to prioritize its text or its purpose, they view the text as uniquely salient—but why this is

so cannot be gleaned from our present experiments. One possibility, supported by pilot data (see Supplementary Analysis 3), is that individuals infer that their peers share a common understanding of the rule's literal meaning, but are more likely to disagree in their appraisals of whether the incident violated the rule's deeper purpose. This perception of the greater univocality of literal meaning could instigate coordination around the rule's text instead of its purpose. However, future studies could directly test this explanation by dissociating salience from text (e.g., rendering the rule's purpose more univocal than its literal meaning).

Though we noted that approaches to statutory interpretation varied substantially across field sites, whether this variation is driven strictly by elements of culture is unclear. Our sampling methods differed across locations, variation in the tendency toward textualism versus purposivism could also partially arise from unobserved differences in the samples' composition.

Conclusion

As part of normative development, adults abandon the uncritical deference to the rule of authority in order to abide by deeper ethical principles (Kohlberg, 1981). Yet, when prompted to decide which behaviors are permissible from the *legal* perspective, this tendency is reversed: People often prioritize the rule's literal scope, even if doing so defies their personal moral values. This textualist tendency was further strengthened by legal training throughout several countries—with legal experts profoundly divorcing their legal and moral assessments of each case. Game theoretic evidence yielded potential insight into the origin of this pattern: Applying a rule's literal meaning, in detriment of its intended purpose or instrumental value, can serve as a focal point (Schelling, 1960) among individuals who knowingly share an interest in aligning their interpretations of the law. In this way, adopting a textualist policy—even while incurring

moral costs in certain instances—could facilitate long-term social coordination (Bennis, Medin, & Bartels, 2010) among lawmakers, citizens, and judges.

Methods

The studies were conducted with approval from the University of Zurich Ethics Committee and Yale's Human Research Protection Program. Study data, analysis scripts, and stimuli (including translations) are available at: https://osf.io/yw8ek/ .

Materials

Our studies employed a battery of nine vignette pairs with one *overinclusion* and one *underinclusion* case in each pair. The coordination game made use of eight vignette pairs (*Vehicles*, *Sleep*, *Driving*, *Library*, *Classroom*, *Shoes*, *Environment*, and *Music*), while the main study employed three pairs (*Classroom*, *Phone*, and *Driving*).

The vignettes first described an incident (e.g., "A 21 year-old woman suffered a traffic accident that took her life. The young woman was driving under the influence."), followed by a description of the rule it gave rise to, including its underlying purpose ("In order to avoid future accidents, Congress passed a zero tolerance policy establishing that: 'If the breathalyzer detects any trace of alcohol, the vehicle will be seized and the driver subject to imprisonment.'"). Then, the vignette described a target act, either in violation of the text of the rule but not its underlying purpose (in overinclusion cases; e.g., using alcohol-based mouthwash prior to driving), or in violation of the purpose of the rule, but not its text (in underinclusion cases; e.g., using ecstasy prior to driving).

Measures

Rule violation judgment. Our main dependent measure was whether the protagonist had violated the rule. In the main study, rule violation judgments (e.g., "Andrea violated the zero-

tolerance policy.") were made on a seven-point scale ranging from 1: 'Strongly Disagree' to 7: 'Strongly Agree'. In the coordination game, rule violation judgments ("Did [the agent] break the rule?") were dichotomous: 1 = 'Yes', 0 = 'No'.

Textualism (difference) score. The mean difference in rule violation judgments between conditions (i.e., overinclusion – underinclusion) constituted our measure of textualism.

Supplementary ratings: literal meaning and moral blame. In the main study, participants in the joint evaluation mode were also asked to rate whether the text of the rule described the target act (e.g., "Andrea drove after ingesting a product containing alcohol."), and whether the protagonist's behavior was morally blameworthy (e.g., 'Andrea is morally blameworthy for what she did.'). Both assessments, i.e., of literal meaning and moral blame respectively, were made on seven-point scales ranging from 1: 'Definitely not' to 7 'Definitely'.

Participants

Laypeople. 3735 participants were recruited in thirteen countries (see Table 2 for demographic information and recruitment details).

Table 2. Sample Composition.

Country	N	Age Mean (SD)	Gender (% women)	Recruitment method
Brazil	207	26.8 (10.0)	51.0%	Word-of-mouth
Colombia	259	22.0 (3.78)	35.4%	Extra-credit
Finland	142	30.3 (13.4)	50.2%	Panel
Germany	359	37.0 (11.4)	50.2%	Panel (www.clickworker.de)
India	254	32.7 (9.50)	63.3%	Panel (www.qualtrics.com)

Italy	350	30.4 (10.9)	50.2%	Panel (www.prolific.co)
Latvia	569	32.7 (9.18)	63.3%	Panel (www.qualtrics.com)
Lithuania	191	32.8 (9.18)	43.0%	Word-of-mouth
Netherlands	391	45.6 (16.7)	48.9%	Panel (www.panelinzicht.nl)
Poland	271	29.5 (11.9)	42.3%	Word-of-mouth
Spain	286	43.4 (15.4)	55.1%	Panel (www.netquest.com)
United Kingdom	202	33.6 (12.7)	57.0%	Panel (www.prolific.co)
United States	254	37.4 (11.2)	57.0%	Panel (www.mturk.com)
Total	3735	36.0 (14.1)	48.5%	-

Legal Experts. As part of the main study, we also recruited 775 lawyers (Age: M = 40.5, SD = 13.9; 48% women) from four countries: Finland (n = 124), the Netherlands (n = 331), Poland (n = 161) and the United States (n = 159).

Coordination Game. 600 participants (Age: M = 26.4, SD = 8.61; 40% women) were recruited via Prolific.co, and invited to take part in an experiment in exchange for monetary compensation.

Procedure: Main Study

In a 2 (Case: overinclusive, underinclusive) × 2 (Evaluation Mode: separate, joint) × 3 (Scenario: Car, Phone, Alcohol) between-subjects design, participants read either an overinclusion or an underinclusion case.

Our primary dependent measure was participants' agreement or disagreement with a statement that the agent had violated the rule. In the joint condition, this question was

accompanied by two additional assessments of the literal scope of the rule and the agent's blameworthiness.

Procedure: Coordination Game

In a 2 between- (Condition: control, game) × 2 within- (Case: overinclusive, underinclusive) × 8 within- (Scenario) balanced incomplete block design, participants read a sequence of six scenarios (plus two filler trials). In the Control condition, participants were asked to "make a decision: Did the person violate the rule (YES) or not (NO)?". Meanwhile, in the Game condition, participants were told:

> You are invited to play the Judging Game. You are Judge 1 and you have been paired with another player, Judge 2. On the following screens, both of you will be reading the same eight stories. Each story describes a rule and a person's behavior. After reading each story, you will both be asked to make a decision: Did the person violate the rule (YES) or not (NO)?
>
> To win extra earnings, you and Judge 2 must agree on as many decisions as possible. You must try and reach the same decision on Case 1, on Case 2, on Case 3, etc., all the way through Case 8 without talking to each other. If you agree on at least six decisions, each of you will earn an additional £1.00 (for a total of £1.70). If not, neither of you will earn the additional £1.00.

Participants made a dichotomous rule violation judgment for each scenario. At the end of the study, participants were randomly paired, and paid a £1 bonus if they agreed on at least six of the eight cases. Study design, predictions and analysis plans were pre-registered at: https://aspredicted.org/blind.php?x=5uv3tj.

References

Barak, A. (2005). *Purposive interpretation in law*. Princeton: Princeton University Press.

Bennis, W. M., Medin, D. L., & Bartels, D. M. (2010). The costs and benefits of calculation and moral rules. *Perspectives on Psychological Science*, *5*(2), 187-202.

Bregant, J., Welbery, I., & Shaw, A. (2019). Crime but not punishment? Children are more lenient toward rule-breaking when the "spirit of the law" is unbroken. *Journal of Experimental Child Psychology*, 178, 266-282.

Bicchieri, C. (2005). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Bostock v. Clayton County, 590 U.S. ___ (2020).

Cushman, F. (2013). Action, outcome, and value: A dual-system framework for morality. *Personality and Social Psychology Review*, *17*(3), 273-292.

Eskridge, W. (1998). Textualism, the Unknown Ideal? *Michigan Law Review* 96.

Fell, J. C., Scherer, M., Thomas, S., & Voas, R. B. (2016). Assessing the impact of twenty underage drinking laws. *Journal of studies on alcohol and drugs*, *77*(2), 249-260.

Fuller, L. (1958). Positivism and fidelity to law: a reply to professor Hart. *Harvard Law Review,* 71(4), 630-672.

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).

Greene, A. S. (2005). The Missing Step of Textualism. *Fordham L. Rev.*, *74*, 1913.

Hart, H. L. A. (1958). Positivism and the Separation of Law and Morals. *Harv. L. Rev.*, *71*, 593.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis*, *15*(3), 199-236.

Kahan, D. M., Hoffman, D., Evans, D., Devins, N., Lucci, E., & Cheng, K. (2015). Ideology or situation sense: an experimental investigation of motivated reasoning and professional judgment. *U. Pa. L. Rev.*, *164*, 349.

Krishnakumar, A. (2022). Cracking the Whole Code Rule, 96 *New York University Law Review.*

Kohlberg, L. (1981). *Essays on Moral Development, Vol. I: The Philosophy of Moral Development*. San Francisco, CA: Harper & Row.

Leal, F. (2015). Six objections to the constitutionalization of private law. *Direitos Fundamentais & Justiça*, n. 33, 123-165.

Manning, J. (2005). Textualism and Legislative Intent, 91 *Virginia Law Review.*

Martínez, E & Tobia, K. (2022). A Survey of the Legal Academy.

Mousavi, S. & Gigerenzer, G. (2017). Heuristics are tools for uncertainty. *Homo Oeconomicus*, 34, 361-379.

Miller, R. M., Hannikainen, I. R., & Cushman, F. A. (2014). Bad actions or bad outcomes? Differentiating affective contributions to the moral condemnation of harm. *Emotion*, *14*(3), 573.

Nucci, L. P., & Turiel, E. (1978). Social interactions and the development of social concepts in preschool children. *Child development*, 400-407.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41-55.

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

Schauer, F. (1988). Formalism. *The Yale Law Journal*, 97(4), 509-548.

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

Sripada, C. S., & Stich, S. (2005). A framework for the psychology of norms. *The innate mind*, *2*, 280-301.

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

Sunstein, C. R. (2005). Moral heuristics. *Behavioral and brain sciences*, *28*(4), 531-541.

Tobia, K. P. (2020). Testing Ordinary Meaning. *Harv. L. Rev.*, *134*, 726.

Tobia, K. P. (2021). Legal Concepts and Legal Expertise, in prep.

Turri, J., & Blouw, P. (2015). Excuse validation: a study in rule-breaking. *Philosophical Studies, 172*(3), 615–634.

Author Contributions

IRH, GA, NS and KPT developed the study concept. IRH and GA analyzed the data. IRH, GA and KPT drafted the first version of the manuscript. All authors were involved in data collection, revised the manuscript and approved the final version for submission.

Funding Information

This research was supported by the Spanish Ministry of Science (PID2020-119791RA-I00; RTI2018-098882-B-I00), the Polish National Science Centre (2020/36/C/HS5/00111; 2017/25/N/HS5/00944), the Swiss National Science Foundation (PZ00P1_179912), and the European Research Council (805498).

Supplementary Analysis 1: Evaluation Modes and Contrast Effects

Previous research found that textualist tendencies were influenced by *evaluation mode* (Struchiner et al., 2020). Judgments were more textualist when participants evaluated both whether the rule was violated *and* whether the agent was moral blameworthy as part of a single task—than when asked to answer the rule violation question in isolation.

To examine whether this effect generalizes to a cross-cultural sample we, probed the two-way interaction between evaluation mode and case-type. In the lay sample, this two-way interaction was significant, albeit very small: Participants were more likely to see overinclusion cases as violations, $B = 0.28$, $t = 2.62$, $p = .009$, but no less likely to see underinclusion cases as violations, $B = -0.14$, $t = -1.37$, $p = .17$. In the lawyer sample, the two-way interaction was not statistically significant. Still, legal experts were more likely to see cases of overinclusion, $B = 0.59$, $t = 2.65$, $p = .008$, but not underinclusion, $B = 0.09$, $t = 0.39$, $p = .69$, as rule violations—when assigned to the joint evaluation condition instead of the separate evaluation condition.

Supplementary Analysis 2: Expertise Effects against Full (Unmatched) Lay Responses

For the sake of completeness, we also compare the lawyer data to the non-matched sample of laypeople. Again, we observed a two-way interaction between expertise and case-type, $F_{(1, 4459)} = 35.88$, $\eta_p^2 = .008$, $p < .001$. Lawyers were more likely to view overinclusion cases as transgressions, $B = 0.37$, $t = 2.68$, $p = .007$, and less likely to view underinclusion cases as transgressions, $B = -0.69$, $t = -5.10$, $p < .001$.

Moderation analyses in the joint evaluation condition revealed no main effect of expertise, $F_{(1, 70)} = 1.16$, $p = .29$, or interaction with literal meaning ratings, $F_{(1, 2210)} = 0.81$, $p = .37$. Once again, an interaction with moral blame ratings was found, $F_{(1, 2236)} = 22.34$, $p < .001$,

η_p^2 = .010. The effect of morality was significant among laypeople, B = 0.15, 95% CI [0.11, 0.19], but not lawyers, B = -0.06, 95% CI [-0.14, 0.02], t = -4.73, p < .001.

Supplementary Analysis 3: Univocality of Literal Meaning and Purpose

In a pilot study, 256 native English speakers were recruited on Prolific.co and asked to consider four rules (*Vehicles*, *Sleep*, *Driving*, and *Library*). For each rule, the participants were asked to consider whether its literal meaning and its purpose would be understood univocally among a group of interpreters (i.e., "Out of 100 people, how many do you think would share the same understanding of the rule's text, what it literally says [/purpose, what it is ultimately for]?"). In a mixed-effects linear regression with participants and scenarios as random effects, we regressed participants'' estimates on a dummy code indicating whether the estimate regarded the literal meaning or the purpose. In this model, the rules' literal meaning was perceived as more univocal (M = 84.2, 95% CI [78.8, 89.5]) than their purpose (M = 78.5, 95% CI [73.2, 83.8]), B = 5.68, t = 7.52, p < .001.