Modeling guilty minds: Laypeople treat knowledge as categorically distinct from recklessness

David G. Kamper[a,b,*], Andrew Bistras[a], Jaime Castrellon[a]

[a] *Department of Psychology, University of California, Los Angeles, Los Angeles, CA, 90095, USA*
[b] *University of Chicago Law School, Chicago, IL, 60637, USA*

Abstract

The Model Penal Code (MPC) distinguishes four levels of criminal culpability (purpose, knowledge, recklessness, and negligence), yet whether laypeople can reliably make these distinctions through genuine inference remains unclear. Prior studies either presented participants with explicitly labeled mental states or asked for self-assessed judgments that may not reflect natural reasoning. We addressed this gap by requiring 212 participants to infer a protagonist's mental state from behavioral evidence in contraband smuggling vignettes with six systematically varied binary features, including the type of contraband and whether payment was offered. Participants classified the protagonist as blameless, negligent, reckless, or knowing and rated the perceived risk of the situation. Cumulative link mixed models showed that risk assessments strongly predicted culpability categorizations and that payment was the only vignette feature to independently shift judgments toward higher culpability after correction for multiple comparisons. Critically, knowledge judgments were associated with lower risk ratings than recklessness judgments, and the proportional odds assumption was violated, with the risk coefficient at the knowledge threshold substantially weaker than at lower thresholds. A features-only Random Forest with SHAP analysis revealed that knowledge was the only category where the type of contraband rather than payment dominated feature importance. Representational

*Corresponding author.
Email address: davidgkamper@ucla.edu (David G. Kamper)

similarity analysis confirmed that semantic distance between vignette texts predicted judgment distance, even though the features driving semantic similarity differed from those driving judgments. These results indicate that laypeople can systematically distinguish MPC mental states through inferential reasoning, but that knowledge functions as a qualitatively distinct category driven by categorical inference rather than extreme risk perception.

Keywords: criminal law, mental states, mens rea, Model Penal Code, culpability judgments, juror decision making

---

1. Introduction

The American criminal justice system assumes that ordinary people can reliably infer the mental states of others from behavioral evidence. Every day, juries reconstruct what a defendant was thinking at the moment of an alleged crime: mental events that are invisible, fleeting, and accessible only through inference. The doctrine of *mens rea* ("guilty mind") holds that criminal culpability requires not merely a prohibited act but also a culpable mental state, reflecting the moral intuition that punishment should be calibrated to blameworthiness (Severance et al., 1992). Whether laypeople can reliably make the graded mental state distinctions that criminal law demands remains an open empirical question with substantial real-world consequences.

1.1. The Model Penal Code framework

Prior to the Model Penal Code (MPC), American criminal law used a patchwork of inconsistently defined terms (willful, wanton, malicious, and so on) to describe culpable mental states (Severance et al., 1992). The MPC, drafted by the American Law Institute and finalized in 1962, replaced this confusion with a systematic four-tier hierarchy of culpability: purpose, knowledge, recklessness, and negligence (American Law Institute, 1985). More than two-thirds of U.S. states have since revised their criminal codes following this framework, making it the dominant structure for assessing criminal culpability in American law.

The four culpable mental states are defined as follows. *Purpose*, the highest culpability level, requires that the prohibited outcome be the actor's "conscious object." *Knowledge* requires awareness that the outcome is "practically certain." *Recklessness* requires conscious disregard of a "substantial and unjustifiable risk." *Negligence*, the lowest culpability level, requires that the actor *should have been* aware of a risk that they failed to perceive. Below negligence lies blamelessness: the absence of any culpable mental state. The hierarchy is inclusive upward, meaning that proof of a higher mental state satisfies a requirement of a lower one (American Law Institute, 1985).

These mental states can be understood as varying along two primary psychological dimensions of risk perception. The first is *risk awareness*: blamelessness involves no awareness of the risk that one's conduct could produce a prohibited outcome, negligence involves a failure to perceive such a risk that should have been perceived, and recklessness and knowledge both involve conscious risk awareness. The second is *risk magnitude*: the reckless actor perceives a substantial risk, while the knowing actor perceives practical certainty. Purpose introduces a third dimension, *volitional orientation*: the purposeful actor not only foresees the outcome but actively desires it as their conscious objective. These distinctions carry enormous legal consequences, often determining the difference between acquittal and decades of imprisonment. Yet they ultimately depend on jurors' ability to make graded inferences about another person's subjective mental states from behavioral evidence alone. The present study focuses on the awareness and risk magnitude dimensions, which differentiate blamelessness, negligence, recklessness, and knowledge. The volitional orientation dimension that distinguishes purpose from knowledge is not tested here because our vignette structure does not support purpose attributions (see Method).

1.2. Prior empirical research

Given the legal stakes involved, little empirical work has examined whether laypeople can make the mental state distinctions the MPC requires. The foundational study by Shen et al. (2011) presented participants with crime scenarios

containing explicitly labeled mental states (e.g., a protagonist who "knew" the package contained drugs) and asked them to classify each scenario into MPC categories. Participants reliably distinguished purpose, negligence, and blamelessness, but confused knowledge and recklessness approximately 45% of the time. This finding raised serious concerns about the psychological validity of the knowledge-recklessness boundary.

Shen's vignettes provided participants with explicit verbal descriptions of mental states; for example, a protagonist described as 'practically certain' or 'aware of substantial risk.' This meant that participants were matching verbal descriptions to legal categories rather than engaging in the inferential reasoning actual jurors perform

However, the Shen methodology contained a fundamental limitation. Participants received explicit verbal descriptions of mental states: for example, a protagonist who is described as "practically certain" or "aware of substantial risk." This meant that the task measured participants' ability to match verbal descriptions to legal categories, something closer to categorization than to the inferential reasoning that actual jurors must engage in. In real legal proceedings, defendants do not make such explicit mental declarations. Jurors must reconstruct these states from testimony about actions, circumstances, and prior behavior.

Addressing this limitation, Vilares et al. (2017) removed explicit labels and asked participants to judge their *own* hypothetical mental states in contraband smuggling scenarios. Using functional magnetic resonance imaging (fMRI) and machine learning, they identified distinct brain activation patterns that classify knowledge versus recklessness. These results provide evidence that judging these mental states evoke separable neurobiological correlates and raise the possibility that these judgments arise from separable cognitive processes. However, two ecological validity concerns remain. First, participants judged their own mental states rather than inferring the mental states of another person, fundamentally altering the task from theory-of-mind inference to introspection (Saxe and Kanwisher, 2003). Second, participants received precise probability information

(e.g., an 80% chance the suitcase contained drugs), eliminating the uncertainty inherent in actual legal scenarios where jurors must estimate a defendant's subjective risk perception from indirect evidence.

1.3. The present study

Together, the Shen and Vilares studies reveal an ecological validity gap. Neither captured the full challenge facing jurors: inferring a third party's mental state from behavioral evidence under realistic uncertainty. We addressed this gap by developing a paradigm in which participants read contraband smuggling vignettes with systematically varied features and inferred the protagonist's mental state without explicit labels or precise probabilities.

Our paradigm focused on the two psychological dimensions that distinguish the three middle tiers of the MPC hierarchy: the risk awareness boundary (between negligence and recklessness) and the risk magnitude boundary (between recklessness and knowledge). We did not include purpose as a response category because the vignette structure (in which a protagonist is recruited by someone else to carry a package) does not support the inference that transporting contraband was the protagonist's conscious objective. Participants instead classified the protagonist into one of four categories spanning the full range from no culpability to high culpability: blamelessness, negligence, recklessness, or knowledge. This design allowed us to test the critical boundaries that prior research has identified as problematic while maintaining ecological validity within a single scenario type.

Prior work on juror decision-making has shown that specific contextual features of criminal scenarios differentially engage the cognitive processes underlying legal judgment, with features tied to social inference exerting stronger effects than features linked to formal legal classification alone (Castrellon et al., 2023). A similar asymmetry may apply to mental state attribution: among the six vignette features in our design, some bear a more direct theoretical relationship to the protagonist's risk awareness; for instance, the type of contraband directly signals whether a prohibited substance is involved, while others, such as

whether payment was offered, may function primarily as aggravating contextual cues that shift judgments without directly informing awareness of criminal risk.

We tested four preregistered hypotheses. First, we predicted that participants' judgments would be distributed non-uniformly across the four categories, indicating systematic rather than random responding (**H1**). Second, we predicted that participants' risk assessments would strongly predict their mental state categorizations, with higher perceived risk associated with higher culpability (**H2**). Third, we predicted that vignette features theoretically relevant to mental state inference would influence judgments more than features with weaker theoretical links to the protagonist's awareness (**H3**). Fourth, we predicted that among aware states, risk magnitude would discriminate knowledge from recklessness, with knowledge trials receiving higher risk ratings (**H4**).

We complemented traditional inferential statistics with two machine learning approaches. Random Forest classification with SHAP (SHapley Additive exPlanations) values provided a nonparametric decomposition of feature importance by mental state category, revealing class-specific patterns invisible to models that assume uniform feature effects across the ordinal scale. Representational similarity analysis (RSA) tested whether the semantic structure of vignette texts corresponds to the structure of participants' judgments, providing a direct test of whether participants respond to surface textual cues or extract abstract inferential features.

2. Method

2.1. Ethics statement

This study was approved by the University of California, Los Angeles Institutional Review Board (IRB-25-0295). All participants provided informed consent prior to participation. The study was preregistered on the Open Science Framework (`https://osf.io/72eg9`).

2.2. Participants

We recruited 230 undergraduate students from the University of California, Los Angeles through the Psychology Department's SONA participant pool. Our target sample size of approximately 200 post-exclusion was determined by a power analysis using effect sizes and design parameters estimated from a 119-participant pilot study. All calculations adjusted for within-participant clustering using the design effect ($\text{DEFF} = 1 + (k-1) \times \text{ICC}$), combining analytic power calculations for H1 and H2 with simulation-based power estimates for the feature effects and the recklessness-knowledge contrast (see pre-registration for full details). At $N = 200$, the study was well-powered for H1 ($> 99\%$), H2 ($> 99\%$), and the payment contrast in H3 (approximately 95%). The source contrast in H3 and the recklessness-knowledge distinction in H4 were underpowered at this sample size and are treated as exploratory. An adaptive minimization algorithm (described below) ensured near-perfect balance across all six binary vignette factors at any stopping point. Eighteen participants were excluded for not completing all five vignettes, yielding a final sample of 212 participants contributing 1,060 trials. Participants' mean age was 20.1 years ($SD = 1.2$, range: 18–24). The sample included 36 freshmen, 47 sophomores, 84 juniors, and 40 seniors (plus 5 post-baccalaureate students). Six participants (2.8%) reported having served on a jury or having other prior legal experience. Participants received course credit for their participation.

2.3. Materials

2.3.1. Vignette design

We developed 64 unique vignette configurations based on the contraband smuggling paradigm of Vilares et al. (2017), which drew on the factual circumstances in *United States v. Jewell* (9th Cir. 1976), a landmark willful blindness case involving contraband transport across the U.S.–Mexico border. In each vignette, a protagonist agrees to transport a package across a national border. Six binary features were systematically varied to manipulate the evidence available for mental state inference.

Source described who asked the protagonist to carry the package: an uncle or a biker from a local bar. *Package* described the container: a gift-wrapped package or a large duffel bag. *Inquiry* described how the protagonist learned about the contents: the source volunteered the information without being asked, or the protagonist actively asked about the contents. *Payment* described whether the protagonist received $1,000 for the transport or no payment. *Location* described the border crossing: Canada or Mexico. *Contents* described what the package actually contained: fentanyl (a potent narcotic) or drugs to treat a rare disease not yet approved in the United States.

These six binary features were fully crossed to generate the 64 unique configurations ($2^6 = 64$). Each participant encountered five vignettes selected through an adaptive trial minimization algorithm. The algorithm tracked the cumulative distribution of vignette configurations selected by all prior participants and used softmax-weighted probabilistic sampling to nudge future selections toward under-represented configurations, keeping the global distribution approximately uniform. It also enforced within-participant diversity by requiring that any two vignettes in a participant's set differ on at least two features. Across the full sample, all features were balanced within 0.8 percentage points of the ideal 50/50 split (range: 49.2% to 50.8%).

The protagonist's name was drawn from a pool of five common male names (John, David, Andrew, Robert, Daniel) and randomized across trials. The specific border city was drawn from a pool of five cities within either Canada (Kingston, Richmond, Thunder Bay, Windsor, Winnipeg) or Mexico (Agua Prieta, Mexicali, Nogales, Nuevo Laredo, Reynosa) and also randomized per trial.

2.3.2. Dependent measures

After reading each vignette, participants provided three responses. First, they classified the protagonist's mental state by selecting one of four descriptions, presented under jury-style instructions asking which mental state the scenario proved "beyond a reasonable doubt." The four response options were shuffled on each trial to prevent order effects. Participants saw only the text

descriptions, not the category labels; the labels used here are for the reader's reference. The four options corresponded to the three MPC culpable mental states tested and a no-culpability baseline. These descriptions were adapted from the MPC § 2.02 definitions (American Law Institute, 1985) and simplified for lay comprehension, following Shen et al. (2011). The three response measures (categorical judgment, risk slider, free-text justification) were always presented in the same order. Whether counterbalancing measure order would alter the relationship between categorical judgments and risk ratings remains an open question for future work.

- *Knowledge*: "[Name] was conscious and aware of the nature of their actions and of the surrounding facts and circumstances."
- *Recklessness*: "[Name] was aware of a substantial and unjustifiable risk of the outcome occurring and consciously ignored that risk."
- *Negligence*: "[Name] failed to use the care that a reasonably careful and thoughtful person would use under similar circumstances."
- *Blamelessness*: "[Name] lacked any of the mental states defined here."

Purpose was not included as a response option because the vignette structure does not support the inference that smuggling was the protagonist's conscious objective. In every scenario, the protagonist is recruited by someone else to carry a package; the question is what the protagonist knew or should have known about the contents, not whether transporting contraband was the protagonist's goal. The four categories presented to participants therefore span the full range from no culpability (blamelessness) through the highest culpability level the scenario can support (knowledge).

Second, participants estimated how likely the protagonist believed that the package contained contraband, using a continuous slider anchored at "[Name] was 100% certain that the package did not contain illicit material" (left endpoint) and "[Name] was 100% certain that the package contained illicit material"

(right endpoint). The slider recorded integer values from 0 to 100 and required participants to move the slider before proceeding. This risk slider provided a continuous measure of the protagonist's perceived risk that could be examined alongside the categorical mental state judgment.

Third, participants provided a free-text justification for their judgment, explaining what contributed to their decision. This open-ended response allowed us to examine whether participants' verbal reasoning aligned with the patterns observed in their structured responses.

2.3.3. Additional measures

Between consecutive vignettes, participants completed one Raven's Progressive Matrices (RPM) item as a brief distractor task. Each participant answered five RPM items total (one after each vignette), drawn from pre-formed five-item booklets randomly assigned to each participant ($M = 3.34$ correct out of 5, $SD = 1.47$).

After completing all five vignettes and their associated RPM items, participants ranked the four mental state definitions from most to least culpable (without seeing the category labels), providing a measure of whether participants understood the intended hierarchy independently of their vignette judgments. Participants also reported their age, educational level, and any prior legal experience.

2.4. Procedure

The study was administered online via Pavlovia using jsPsych. After providing informed consent, participants read instructions framing the task as jury service for five separate criminal cases, each involving a defendant charged with "unlawful trafficking" (transporting an illegal item across an international border). Participants were told that none of the cases were identical and were asked to read each scenario closely. They then completed five vignette trials in randomized order. On each trial, participants (a) read the vignette text and selected a mental state classification, (b) rated the protagonist's perceived risk on

the continuous slider, and (c) wrote a free-text justification, each on a separate screen. A single RPM item appeared after each vignette as a brief distractor. Following the fifth trial, participants completed the MPC ranking task and provided demographic information. The entire session lasted approximately 20 minutes. See Fig. 1 for an example trial.

2.5. Analytical approach

2.5.1. Primary analysis: Cumulative link mixed models

We employed cumulative link mixed models (CLMMs) as the primary analytical framework, implemented via the `ordinal` package (Christensen, 2019) in R (**?**). CLMMs are the appropriate model for ordered categorical outcomes with clustered observations. Unlike standard ordinal logistic regression, CLMMs include random intercepts that account for the within-participant clustering of five trials per person. Treating the 1,060 trials as independent observations would inflate the effective sample size and produce artificially narrow confidence intervals. The intraclass correlation for the culpability variable was 0.162 (design effect = 1.65), confirming that the clustering adjustment was necessary.

All CLMMs used a logit link and flexible (estimated) thresholds. For hypothesis testing, we fit three nested models: a null model with random intercepts only, a risk-only model adding the slider value as a fixed effect (H2), and a full model adding the six binary vignette features (H3). Model comparison used likelihood ratio tests. For the knowledge-recklessness discrimination (H4), we fit a linear mixed model predicting slider values from a binary knowledge indicator among trials classified as either recklessness or knowledge, with a random intercept for participant.

To test the proportional odds assumption, we fit separate binary logistic mixed models at each threshold of the ordinal variable (blamelessness vs. higher, negligence-or-below vs. higher, recklessness-or-below vs. knowledge) and compared the risk slider coefficient across thresholds. We also computed ROC curves and AUC values for each threshold-specific model to characterize discriminative accuracy.



Figure 1: Example trial from the experiment. Each trial consisted of three screens presented in sequence. *Top*: After reading the vignette, participants selected one of four mental state descriptions under jury-style instructions. The four response options were shuffled on each trial. *Bottom*: Participants rated the protagonist's perceived risk of carrying contraband on a continuous 0–100 slider. On a third screen (not shown), participants provided a free-text justification for their judgments.

2.5.2. Complementary analyses: Random Forest, SHAP, and RSA

We complemented the CLMMs with three machine learning analyses, implemented in Python using scikit-learn (Pedregosa et al., 2011). First, we trained Random Forest classifiers (Breiman, 2001) to predict the four-level culpability judgment. To avoid the circularity of including a participant's own risk rating as a predictor of their own judgment, we trained two separate models: a features-only model using the six binary vignette features, and a features-plus-slider model adding the risk rating. Both used 500 trees, balanced class weights, and 10-fold cross-validation stratified by judgment category. Folds were constructed at the trial level, so that trials from the same participant could appear in different folds. For the features-only model this does not introduce data leakage, because the six predictors are properties of the vignette, not of the participant; knowing one participant's feature values on one trial provides no information about a different trial's features. For the features-plus-slider model, within-participant similarity in slider usage could modestly inflate accuracy estimates. We retain the trial-level split for comparability across models, and we note that the slider model serves only as a descriptive benchmark — no inferential claims rest on its accuracy.

Second, we computed SHAP (SHapley Additive exPlanations) values (Lundberg and Lee, 2017) from the features-only Random Forest to decompose feature importance by mental state category. SHAP values quantify each feature's contribution to the predicted probability of each class for each observation. Because CLMMs assume a single set of feature effects across all thresholds, they cannot detect patterns where different features matter for different categories, whereas SHAP values can.

Third, we conducted representational similarity analysis (RSA) to test whether the semantic structure of the vignette texts corresponds to the structure of participants' culpability judgments. We encoded the 64 unique vignette texts using two sentence-transformer models (Reimers and Gurevych, 2019) with different architectures: BAAI/bge-large-en-v1.5 (1,024 dimensions, CLS-token pooling,

trained with hard-negative contrastive learning) and all-MiniLM-L6-v2 (384 dimensions, mean pooling). We used two models because our vignettes are generated from a shared template with only small word-level changes across conditions, and mean-pooling models can produce near-identical embeddings for texts that share most of their tokens. The contrastive-trained BGE model addresses this concern. We also conducted a clause-level RSA in which we embedded only the text fragments that vary between conditions (e.g., "a biker" vs. "his uncle"), eliminating shared template text entirely. For each analysis, we computed pairwise cosine distances between vignette embeddings and correlated these with pairwise differences in mean culpability scores using Spearman's ρ.

3. Results

3.1. Judgments are non-uniform across mental state categories (H1)

Participants' mental state judgments were distributed non-uniformly across the four categories ($\chi^2(3) = 131.67$, $p < .001$, Cohen's $w = 0.453$; chi-square statistic divided by the design effect of 1.65 to correct for within-participant clustering). Negligence was the modal response (452 trials, 42.6%), followed by recklessness (286, 27.0%), blamelessness (177, 16.7%), and knowledge (145, 13.7%). The predominance of negligence is notable: participants overwhelmingly characterized the protagonist as someone who *should have known* rather than someone who *did know*. Fig. 2 displays the distribution against a uniform baseline. Presentation order had no significant main effect on judgments ($b = -0.007$, $z = -0.17$, $p = .864$), and no feature $\times$ position interaction survived Holm correction for six comparisons (smallest $p_{\text{Holm}} = .241$; see Table A.2).

3.2. Risk perception strongly predicts mental state categorization (H2)

Participants' risk slider ratings were strongly associated with their mental state categorizations (Spearman $\rho = .480$, $p < .001$). A CLMM with slider value as the sole predictor confirmed that perceived risk robustly predicted ordinal

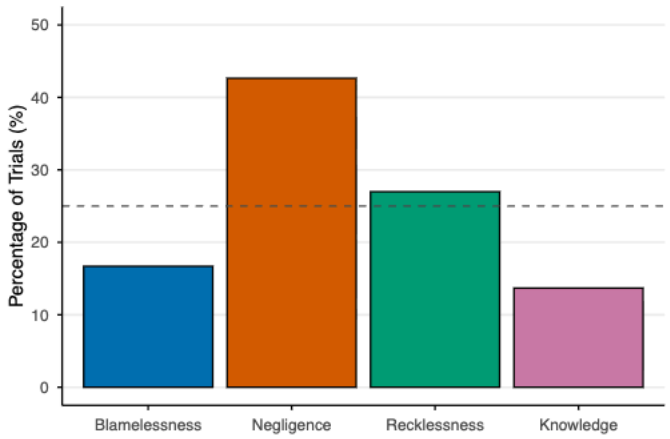


Figure 2: Distribution of mental state judgments across 1,060 trials from 212 participants. The dashed line indicates the expected frequency under a uniform distribution ($n = 265$ per category). Negligence was selected on 42.6% of trials, more than three times as often as knowledge (13.7%).

culpability ($b = 0.046$, $SE = 0.003$, $z = 14.87$, $p < .001$). Each 10-point increase on the slider multiplied the cumulative odds of a higher culpability judgment by 1.58. The participant-level random intercept variance was 0.87 ($SD = 0.93$), indicating substantial individual differences in overall culpability tendency. Full model coefficients for both the risk-only and the full model (with vignette features) are reported in Table 1.

Fig. 3 displays risk slider distributions by judgment category. Mean risk ratings increased monotonically from blamelessness through recklessness ($M_{\text{B}} = 21.7$, $M_{\text{N}} = 34.3$, $M_{\text{R}} = 59.1$), but knowledge ratings ($M_{\text{K}} = 51.4$) fell *below* recklessness, a pattern we investigate in the H4 analysis.

Fig. 4 shows predicted probabilities from the CLMM as a function of risk slider value. Negligence is the most probable category across a wide range of the slider (approximately 0–50), peaking near a slider value of 25. At the lowest risk values, blamelessness is nearly as probable as negligence, but it declines steeply as risk increases. Recklessness rises with the slider and becomes the most probable category above approximately 60. The knowledge curve increases

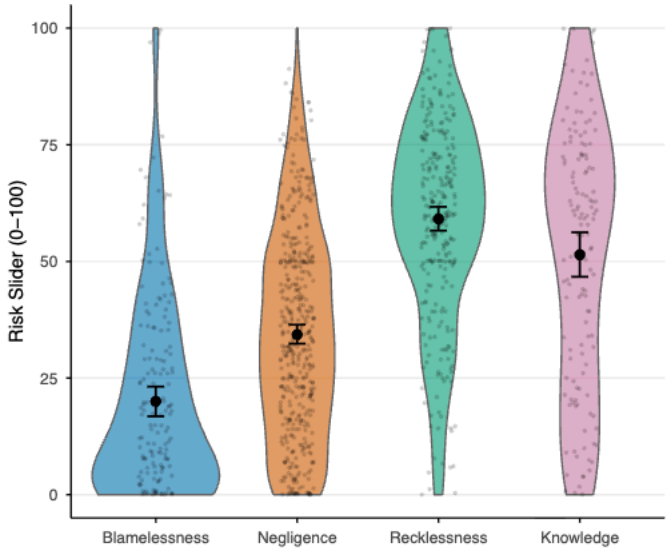


Figure 3: Risk slider ratings (0 to 100) by mental state judgment. Points represent individual trials; black dots and error bars show means ± 95% confidence intervals. Risk ratings increase monotonically from blamelessness through recklessness but decrease for knowledge, indicating a reversal at the top of the hierarchy.

gradually and exceeds recklessness only at the extreme upper end of the scale (above approximately 90), where the confidence bands for the two categories overlap substantially. This pattern foreshadows the proportional odds violation described below: risk perception discriminates well among the lower three categories but provides little separation between recklessness and knowledge.

3.3. Payment is the only vignette feature to survive correction (H3)

Adding the six binary vignette features to the CLMM significantly improved model fit over the risk-only model (likelihood ratio $\chi^2(6) = 23.58$, $p < .001$). Table 1 reports the full model coefficients. The improvement was driven primarily by a single feature: payment. Receiving \$1,000 shifted judgments substantially toward higher culpability ($b = 0.53$, $z = 3.91$, $OR = 1.70$, $d \approx 0.29$), and this was the only feature to survive Holm correction for six comparisons ($p_{\text{Holm}} < .001$). Source (biker vs. uncle) was in the expected direction but was

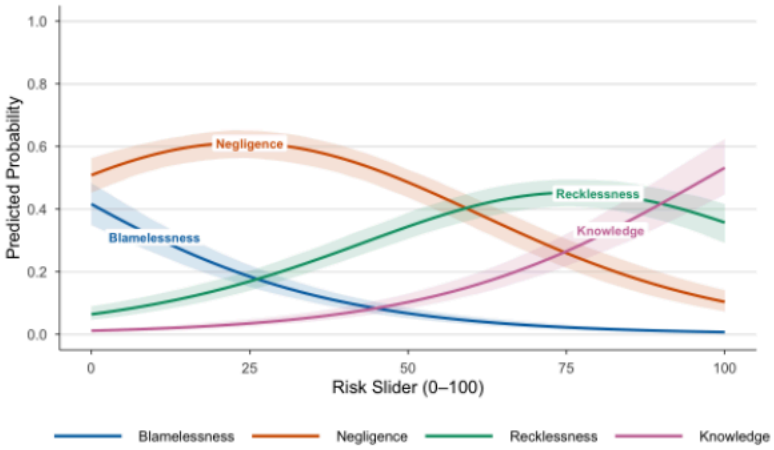


Figure 4: Predicted probability of each mental state category as a function of risk slider value, derived from the risk-only CLMM. Shaded regions indicate 95% confidence bands. Negligence dominates across slider values of approximately 0–50. Recklessness becomes the most probable category above approximately 60. The knowledge curve rises gradually and exceeds recklessness only at the extreme upper end of the scale, where the confidence bands overlap substantially, reflecting the weak discriminative power of risk perception at the top of the culpability hierarchy.

not statistically significant ($b = 0.25$, $z = 1.90$, $p = .057$). Contents (fentanyl vs. rare disease medication) was also not statistically significant ($b = -0.23$, $z = -1.84$, $p = .065$), though the direction was unexpected, with fentanyl associated with *lower* rather than higher culpability, a pattern we return to in the SHAP analysis below. The remaining three features (location, package, and inquiry) had no significant effects. Fig. 5 displays the coefficient estimates with 95% confidence intervals.

The pattern of significance carries theoretical weight. Payment signals that the protagonist had a financial motive and reason to suspect illicit contents. It is the feature most directly diagnostic of the protagonist's awareness and motivation. Although the effect of source was in the expected direction, with a stranger (biker) increasing culpability relative to a family member (uncle), it was not statistically significant. The features with weaker theoretical links to the

Table 1: Cumulative link mixed model coefficients (N = 1,060 trials from 212 participants).

	Model 1: Risk only			Model 2: Risk + features			
Predictor	b	SE	z	b	SE	z	p_{Holm}
Risk slider	0.046	0.003	14.87***	0.041	0.003	12.88***	
Payment ($1000)				0.531	0.136	3.91***	<.001
Source (biker)				0.247	0.130	1.90	.285
Contents (fentanyl)				−0.229	0.124	−1.84	.261
Inquiry (source tells)				−0.179	0.124	−1.44	.474
Package (duffel bag)				0.090	0.125	0.72	.934
Location (Mexico)				0.010	0.124	0.08	.934
Threshold coefficients							
$\alpha_{1\|2}$ (B|N)	−0.339			−0.320			
$\alpha_{2\|3}$ (N|R)	2.504			2.582			
$\alpha_{3\|4}$ (R|K)	4.457			4.564			
Random effects							
Participant σ^2 (SD)	0.872 (0.934)			0.908 (0.953)			
Model fit							
Log-likelihood	−1200.48			−1188.69			
AIC	2410.96			2399.38			
LR test (Model 1 vs. 2): $\chi^2(6) = 23.58$, $p < .001$							

Note. *** $p < .001$. Positive coefficients indicate higher culpability judgments. p_{Holm} = Holm-corrected p-value for the six vignette features (risk slider not included in correction). B = Blamelessness; N = Negligence; R = Recklessness; K = Knowledge.

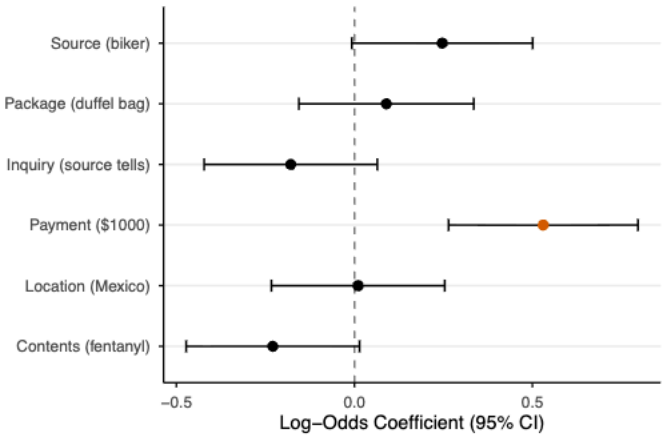


Figure 5: Log-odds coefficients (with 95% confidence intervals) for the six binary vignette features from the full CLMM. Positive values indicate that the feature shifts judgments toward higher culpability. Payment is the only feature whose confidence interval excludes zero after Holm correction.

protagonist's mental state (location, package type, and whether the protagonist asked about the contents) had no significant effects. This pattern suggests that participants' attributions were driven by features diagnostic of the protagonist's mental state rather than by geographic stereotypes or surface characteristics of the scenario.

3.4. Knowledge judgments are associated with lower risk ratings than recklessness (H4)

We predicted that knowledge judgments would be associated with higher risk ratings than recklessness judgments, reflecting the MPC's definition of knowledge as "practical certainty." The data revealed the opposite. Among the 431 trials classified as either recklessness or knowledge, knowledge trials received significantly *lower* risk slider ratings ($M = 51.4$, $SD = 28.7$) than recklessness trials ($M = 59.1$, $SD = 22.3$; mean difference = 6.96 slider points, $t = -3.06$, $p = .002$, $d = -0.28$).

This reversal suggests that knowledge is not simply "high recklessness" in participants' mental models. Instead, knowledge judgments appear to be trig-

gered by qualitative features of the scenario, cues suggesting that the protagonist *must have known*, rather than by extreme values on a continuous risk dimension. We tested this interpretation with two converging analyses.

3.4.1. The proportional odds assumption is violated

If the four mental states formed a single ordinal continuum driven by risk perception, the CLMM's proportional odds assumption would hold: the risk coefficient would be approximately constant across all thresholds. Standard proportional odds tests do not accommodate random effects, so we compared threshold-specific coefficients from separate binary logistic mixed models. The risk coefficient declined monotonically: $b = 0.059$ at the blamelessness-negligence transition, $b = 0.052$ at negligence-recklessness, and $b = 0.019$ at recklessness-knowledge. The coefficient at the knowledge threshold was 3.2 times smaller than at the blamelessness threshold, and the 95% confidence intervals for the lowest and highest thresholds did not overlap. ROC analysis confirmed this pattern: discriminative accuracy was excellent at the blamelessness-negligence boundary (AUC = 0.926), strong at negligence-recklessness (AUC = 0.877), and weakest at recklessness-knowledge (AUC = 0.859). Risk perception is a powerful discriminator among the lower three categories but a weak predictor of the transition to knowledge. Fig. 6 displays the threshold-specific coefficients.

3.4.2. SHAP analysis reveals class-specific feature importance

To examine whether different features drive different mental state attributions, we trained Random Forest classifiers on three predictor sets: (a) the six binary vignette features alone, (b) the risk slider alone, and (c) the six features plus the risk slider. All models used 500 trees, balanced class weights, and 10-fold stratified cross-validation. The features-only model achieved modest accuracy (30.7% $\pm$ 3.5%, $\kappa = 0.074$), above the 25% uniform chance baseline but below the 42.6% majority-class rate. The risk-only model substantially outperformed the features-only model (39.9% $\pm$ 3.5%, $\kappa = 0.189$; paired $t(9) = 5.86$, $p < .001$), confirming that subjective risk perception captures more judgment-

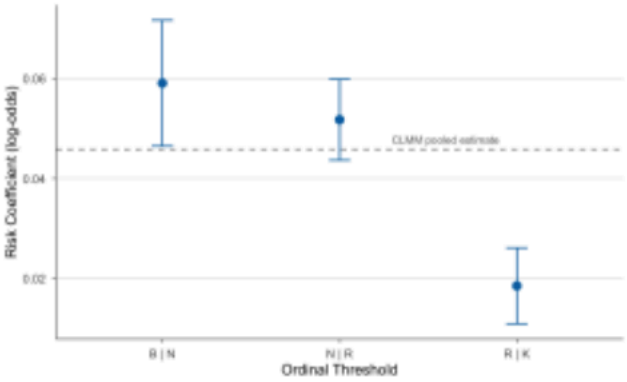


Figure 6: Risk slider coefficient (log-odds) estimated separately at each ordinal threshold. The dashed line shows the pooled CLMM estimate. The coefficient at the knowledge threshold ($b = 0.019$) is 3.2 times smaller than at the blamelessness threshold ($b = 0.059$), with non-overlapping confidence intervals.

relevant variance than objective stimulus features. The combined model significantly outperformed both individual models (47.3% $\pm$ 2.9%, $\kappa = 0.254$; vs. features-only: paired $t(9) = 11.15$, $p < .001$; vs. risk-only: paired $t(9) = 6.82$, $p < .001$). The fact that the combined model exceeds risk-only indicates that vignette features carry predictive signal that is not fully mediated by participants' risk assessments.

We computed SHAP values from the features-only Random Forest to decompose feature importance by mental state category. The class-specific analysis revealed distinct importance profiles across the four categories (Fig. 7). For blamelessness, source (biker vs. uncle) was the dominant feature (mean |SHAP| = 0.115), with payment second (0.070). For negligence and recklessness, payment was the most important feature (0.057 and 0.088, respectively). For knowledge, however, a qualitatively different pattern emerged: contents (fentanyl vs. rare disease medication) became the most important feature (0.049), surpassing payment (0.042). Knowledge was the only category where the type of contraband mattered more than the financial incentive.

This class-specific pattern helps explain the unexpected negative contents

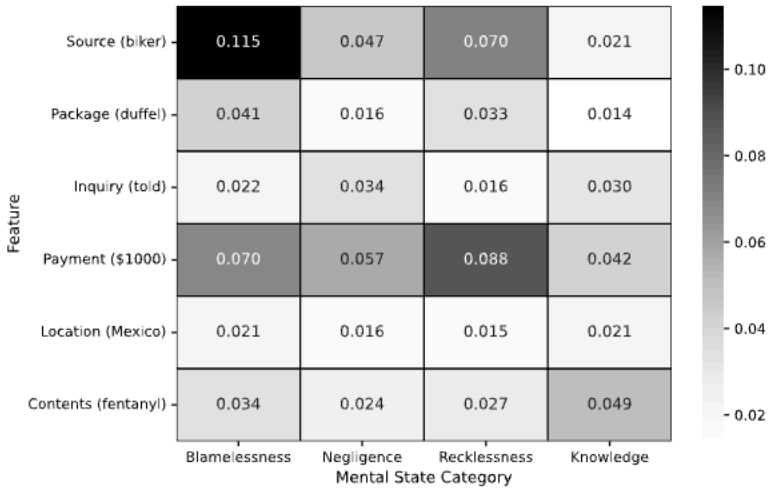


Figure 7: Mean absolute SHAP values by feature and mental state category from the features-only Random Forest classifier. Darker cells indicate greater feature importance. Source dominates blamelessness attributions. Payment dominates negligence and recklessness. For knowledge, contents (fentanyl) emerges as the most important predictor, the only category where the type of contraband outweighs financial incentive.

coefficient in the CLMM ($b = -0.23$). Fentanyl predominantly appears in trials judged as blamelessness, negligence, or recklessness, where its pooled SHAP effect is negative. Yet it is also the feature that most distinguishes the minority of knowledge attributions. Because the CLMM estimates a single pooled effect, the negative majority-category effect masks fentanyl's outsized role in knowledge specifically. This pattern suggests that participants inferred knowledge from what was in the package—a categorical, "must have known" cue—rather than from the same risk-related features that drove recklessness attributions.

3.5. Semantic similarity predicts judgment similarity, but not through the expected features

Representational similarity analysis tested whether vignettes that are closer in semantic embedding space elicit more similar culpability judgments. The results depended critically on the embedding model and analysis level (Fig. 8).

Using BGE-large (a contrastive-trained model with CLS-token pooling), semantic distance significantly predicted judgment distance ($\rho = 0.077$, $p < .001$, $N = 2{,}016$ vignette pairs). Using MiniLM (a mean-pooling model), the same analysis yielded a null result ($\rho = 0.019$, $p = .395$). The difference between models arises because our vignettes share approximately 90% of their text (the template), and mean-pooling architectures produce near-identical embeddings when most tokens are shared, diluting the signal from the few words that actually vary between conditions. The contrastive-trained BGE model, optimized to distinguish similar texts via CLS-token pooling, resolves this dilution.

To further confirm this interpretation, we conducted a clause-level RSA in which we embedded only the text fragments that varied between conditions (e.g., "a biker" vs. "his uncle," "fentanyl" vs. "drugs to treat a rare disease"). This analysis produced the strongest effect ($\rho = 0.129$, $p < .001$), consistent with the hypothesis that the shared template text was attenuating the full-text correlations.

Importantly, the features that produced the largest semantic distances were not the features that most influenced judgments. Location (Canada vs. Mexico) and contents (fentanyl vs. rare disease medication) created the largest embedding shifts (mean cosine distances of 0.044 and 0.038, respectively), while payment, the strongest behavioral predictor, produced one of the smallest shifts (0.012). This dissociation is not simply an artifact of how numbers are embedded: the payment manipulation is expressed as a full clause ("agrees to bring back the package for $1000" vs. "agrees to bring back the package without asking for payment"), and the small cosine distance reflects the high lexical overlap between the two versions rather than the proximity of numerical tokens. Participants' mental state attributions were organized by a feature that is semantically subtle (whether money changed hands) but inferentially diagnostic, rather than by features that are semantically prominent but less relevant to the protagonist's mental state. This dissociation provides evidence that participants engaged in genuine inferential reasoning rather than responding to surface-level textual cues.

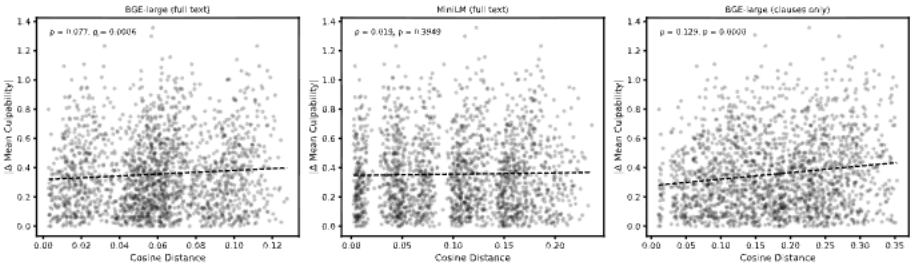


Figure 8: Representational similarity analysis across three analytical approaches. Each point represents one pair of the 64 unique vignette configurations (N = 2,016 pairs). Left: BGE-large full-text embeddings show a significant positive relationship between cosine distance and judgment distance. Center: MiniLM full-text embeddings show no relationship, demonstrating the template dilution problem with mean-pooling architectures. Right: BGE clause-level embeddings (varying text only) show the strongest effect, confirming that the shared template text attenuates the full-text correlation.

3.6. Participants understand the culpability hierarchy

Independently of the vignette task, 40.6% of participants ranked the four mental state definitions in the correct order (knowledge > recklessness > negligence > blamelessness), far exceeding the 4.17% expected by chance based on 4!=24 possible orderings ($p < .001$, binomial test). The mean Kendall's τ between participants' rankings and the correct ordering was 0.623 ($SD = 0.504$), significantly above zero ($t(211) = 17.99$, $p < .001$). These results indicate that most participants understood the intended hierarchy, and that the vignette-based judgment patterns described above cannot be attributed to confusion about what the categories mean.

3.7. Additional analyses

Cognitive ability, as measured by Raven's Progressive Matrices performance, did not moderate the relationship between risk perception and culpability judgment (RPM × risk slider interaction: $z = -0.04$, $p = .965$). This null interaction indicates that the risk-to-judgment mapping operated similarly across the range of cognitive ability observed in our sample.

Free-text justification analysis provided convergent evidence. A multinomial logistic regression classifier trained on BGE sentence-transformer embeddings of participants' justifications predicted mental state judgments at 51.9% accuracy (10-fold cross-validated, $\kappa = 0.24$), well above the 25% chance baseline and above the 42.6% majority-class rate. We used multinomial rather than ordinal classification here because the goal was predictive accuracy with high-dimensional embeddings (1,024 features) rather than inference about ordinal structure. A model combining vignette features with justification embeddings significantly outperformed both vignette features alone (paired $t(9) = 13.99$, $p < .001$) and vignette features plus risk slider (paired $t(9) = 3.83$, $p = .004$). These results demonstrate that participants' verbal reasoning carries predictive signal beyond both the objective stimulus features and their own risk ratings, indicating that the structured measures do not fully capture the information participants used when making their judgments. BERTopic analysis of the justification texts identified 23 interpretable topics, with several showing differential associations with specific mental state categories (e.g., payment-related reasoning associated with recklessness, trust-related reasoning with negligence, certainty-related language with knowledge; see Supplementary Materials for full topic descriptions and distributions).

4. Discussion

We examined whether laypeople can distinguish the Model Penal Code's mental state categories when required to infer a protagonist's mental state from behavioral evidence under realistic uncertainty. Three main findings emerged.

4.1. Laypeople make systematic mental state distinctions

Our most basic finding is that participants' judgments were far from random. The distribution of mental state attributions deviated strongly from uniformity ($w = 0.453$), risk perception predicted culpability categorization with a large effect ($z = 14.87$), and the addition of vignette features significantly improved

predictive accuracy beyond risk perception alone. These results extend the findings of Shen et al. (2011) and Vilares et al. (2017) by demonstrating that laypeople respond systematically even when mental states are neither labeled nor self-assessed, under conditions that approximate the inferential challenge faced by actual jurors. Participants drew boundaries around psychologically meaningful features (payment, source) rather than surface characteristics (location, package type), consistent with genuine mental state reasoning rather than superficial categorization.

A three-way Random Forest comparison clarified the relationship between objective stimulus features and subjective risk perception. Risk perception alone (39.9%) substantially outperformed the six vignette features alone (30.7%), confirming that subjective risk assessment is the more proximal driver of judgments. However, the combined model (47.3%) significantly outperformed risk alone (paired $t(9) = 6.82$, $p < .001$), indicating that vignette features carry predictive signal that is not fully mediated by participants' risk assessments. Features appear to influence judgments both indirectly, by shaping risk perceptions, and through independent pathways.

Payment was the most influential feature and the strongest theoretical indicator of awareness and motivation: receiving money to transport a package provides a reason to suspect illicit contents. Payment was the only robust predictor after correction for multiple comparisons. The effect of source (stranger vs. family member) was in the expected direction but was not statistically significant. Features with weaker theoretical links to the protagonist's mental state, including geographic location and package type, had no significant effects. This pattern supports the construct validity of participants' mental state attributions: they responded to features that provide information about the protagonist's state of mind rather than to surface characteristics of the scenario.

4.2. Knowledge is qualitatively different from recklessness

The most consequential finding is the knowledge-recklessness reversal. The MPC defines knowledge as a higher culpability state than recklessness, implying

that knowledge should be associated with greater perceived risk. Our participants showed the opposite: knowledge judgments were associated with lower risk ratings than recklessness judgments ($d = -0.28$). This reversal is unlikely to reflect noise from a small or heterogeneous category: the effect held in a mixed model with participant random intercepts (controlling for individual differences in overall tendency), and the SHAP analysis showed a qualitatively different feature importance profile for knowledge as a class, not merely a noisier version of the recklessness profile. Four converging lines of evidence support the conclusion that the reversal reflects a genuine qualitative difference.

First, the direct comparison was significant ($p = .002$) with a sample of 431 trials. Second, the proportional odds analysis revealed that the risk coefficient at the knowledge threshold was 3.2 times smaller than at the blamelessness threshold, with non-overlapping confidence intervals. Third, ROC analysis showed that discriminative accuracy declined monotonically across thresholds, with the risk slider least able to distinguish recklessness from knowledge. Fourth, the SHAP analysis showed that knowledge was the only mental state category where the type of contraband (contents) was a more important predictor than payment. For blamelessness, negligence, and recklessness, the feature importance profiles were organized around payment and source, features related to the protagonist's motivation and social context. For knowledge, the dominant feature was what was actually in the package.

Analysis of judgment distributions under progressively stricter awareness conditions (Table A.5) further supports this interpretation. When trials were restricted to those combining the strongest awareness cues (payment, biker source, and active inquiry), recklessness replaced negligence as the modal response, consistent with participants moving along a continuous risk dimension as awareness-relevant evidence accumulated. Yet knowledge attributions remained at approximately 16% even under the most extreme conditions, suggesting that accumulating risk-relevant features shifts participants from negligence toward recklessness but does not push them into knowledge. The transition to knowledge appears to require a qualitatively different inference.

We interpret this convergence as evidence that knowledge, in participants' intuitive psychology, is not simply "extreme recklessness" but a qualitatively distinct category. When participants attributed knowledge, they appear to have reasoned that the protagonist *must have known* based on the nature of the situation rather than concluding that the protagonist perceived an extremely high probability of risk. The presence of fentanyl, a well-known and dangerous narcotic, may have served as a categorical signal: if the package contains something this obviously illegal, the carrier must have known. This inference may not require an estimate of subjective probability. Instead, it may rest on a judgment about whether the evidence, taken as a whole, implies that the protagonist knew what they were carrying.

This interpretation aligns with legal scholarship on willful blindness, where courts have recognized that knowledge can sometimes be inferred from the totality of circumstances rather than from direct evidence about a defendant's subjective probability estimates (Global-Tech Appliances, Inc. v. SEB S.A., 2011). The factual scenario underlying our vignettes, drawn from the circumstances in *United States v. Jewell* (9th Cir. 1976), was itself a case in which the court held that deliberate avoidance of knowledge could satisfy the knowledge requirement. Our participants appear to have applied a similar inferential strategy: treating knowledge as something that can be concluded from situational features rather than measured on a probability scale. This pattern is also consistent with recent neuroscience evidence that jurors rely more heavily on narrative-based inference than on probability estimation when evaluating evidence in criminal cases (Castrellon et al., 2022).

This finding also reframes the knowledge-recklessness confusion reported by Shen et al. (2011). Those authors attributed the confusion to a failure of psychological discrimination. Our results suggest an alternative: the boundary between knowledge and recklessness may be genuinely qualitative rather than quantitative. Presenting the distinction as a probability threshold ("practically certain" vs. "substantial risk") may obscure the categorical inference that actually drives knowledge attributions. When participants must infer mental states from be-

havior rather than match labels to definitions, the knowledge-recklessness distinction appears categorical rather than continuous. The boundary depends on whether the evidence implies that the protagonist knew the nature of what they were carrying, not on the height of their subjective probability estimate.

4.3. Judgments reflect inferential reasoning, not surface textual similarity

The RSA results contribute both a substantive finding and a methodological lesson. Substantively, the features that produced the largest semantic distances in the embedding space (contents, location) had the smallest effects on judgments, while the feature with the smallest semantic footprint (payment) had the largest effect. This dissociation rules out a "textual salience" account of mental state attribution: participants were not simply responding to whichever words stood out most in the vignette text. They extracted abstract, inferentially relevant features from the scenarios.

Methodologically, the comparison between embedding models demonstrates that the RSA results in factorial vignette designs are sensitive to model architecture. Our vignettes, like those in many experimental paradigms, are generated from a shared template with small word-level changes across conditions. Mean-pooling models produce near-identical embeddings for such texts because the shared template tokens dominate the average. The contrastive-trained BGE model, which uses CLS-token pooling optimized to distinguish similar texts, resolved this dilution and revealed a significant relationship. The clause-level analysis, which embedded only the varying text fragments, produced the strongest effect. Researchers using RSA with templated stimuli should consider both the embedding architecture and the possibility of analyzing varying content separately from shared template text.

4.4. Limitations

There are some limitations that warrant acknowledgment. Our sample consisted of undergraduate students at a single university, a population that may

not represent the diversity of actual jury pools in terms of age, education, socioeconomic background, or legal knowledge (Henrich et al., 2010). Although 40.6% of participants correctly ranked the mental state categories (far above chance), a substantial minority did not, and the degree to which our findings generalize to more representative samples remains to be tested.

The vignette paradigm used a single scenario type (contraband smuggling across a national border). This scenario was well suited to testing the awareness and risk magnitude boundaries of the MPC hierarchy, but the degree to which our findings generalize across crime types is unknown. The contraband context may be particularly likely to elicit categorical knowledge inferences ("either you knew it was fentanyl or you did not"), and other crime types involving more continuous risk dimensions might produce different patterns. Scenarios involving, for example, financial fraud or environmental harm could test whether the qualitative nature of knowledge attributions is specific to contraband or reflects a general feature of how people reason about knowledge.

Moreover, the smuggling paradigm positions the protagonist as someone recruited to transport a package rather than as the instigator of the criminal act. This accomplice-like framing may interact with how participants reason about knowledge. Girgis (2013) has argued that payment for participation in a criminal enterprise can signal knowing involvement rather than mere risk-taking, a framework that aligns with the strong payment effect we observed. The knowledge-recklessness boundary may therefore be context-dependent: more categorical in accomplice contexts like ours, where the judgment shifts from "how probable did they think it was?" to "did they knowingly sign onto the criminal plan?," but more gradient in perpetrator contexts like those studied by Shen et al. (2011). Importantly, participants were never instructed to reason about accomplice liability. Their task was to evaluate what the protagonist knew and how culpable they were: a mental state attribution process that is not specific to accomplice contexts. Future work varying the protagonist's role in the offense could test the context sensitivity of mental state attributions.

Our paradigm did not include purpose as a response category because the

vignette structure does not support purpose attributions. The boundary between purpose and knowledge, which involves volitional orientation (whether the outcome was the actor's conscious objective), remains untested. Future work using scenarios in which the protagonist initiates the criminal act could examine whether the purpose-knowledge distinction also maps onto a qualitative rather than quantitative psychological difference.

Each participant saw only 5 of the 64 possible vignette configurations. The adaptive trial minimization algorithm ensured that these 5 trials were diverse (differing on at least two features from one another) and that the global distribution of configurations across all participants was approximately uniform. However, 5 observations per person may not be sufficient to estimate individual-level feature sensitivities. The CLMM random intercepts capture individual differences in overall culpability tendency, but the random slopes model (which would estimate per-participant risk sensitivity) did not converge. Designs with more trials per participant would permit more robust within-person analyses.

Finally, the unexpected negative effect of contents (fentanyl associated with lower culpability) deserves further investigation. One possibility is that the participants viewed fentanyl as so dangerous that they judged the protagonist unlikely to have knowingly transported it, inferring that no one would willingly carry a substance with such severe legal consequences. This would represent a form of reverse inference ("it is too dangerous, so they must not have known") that runs counter to the straightforward prediction that more dangerous contents should increase perceived culpability. Testing this interpretation would require follow-up studies that manipulate the severity of the consequences more directly.

4.5. Implications for legal practice

If the pattern we observe generalizes, it has practical implications for how the knowledge-recklessness distinction is communicated to jurors. Jury instructions that frame this distinction purely in probabilistic terms ("practically certain" vs. "substantial risk") may not capture how jurors actually make the judgment.

Our data suggest that jurors may rely on categorical inferences about what the defendant must have known given the circumstances, rather than comparing subjective probability estimates to a threshold. Instructions that describe knowledge in terms of circumstantial evidence ("the nature of the situation was such that the defendant must have been aware") rather than in terms of probability ("the defendant was practically certain") might better align with how jurors reason.

The strong payment effect also has practical relevance. Evidence that a defendant received financial compensation for their role in a criminal act was the single most powerful feature shifting mental state attributions toward higher culpability.

4.6. Conclusion

Laypeople can systematically distinguish Model Penal Code mental states when required to infer those states from behavioral evidence under realistic uncertainty. Their judgments are driven primarily by their own risk assessments and by whether the protagonist received payment, while features with weaker links to the protagonist's mental state have negligible influence. The knowledge category, however, operates differently from the lower three categories. It is driven by qualitative, categorical inferences about what the protagonist must have known rather than by extreme values on a continuous risk dimension. The MPC's mental state hierarchy has psychological validity as a framework for how people make culpability judgments, but the boundary between knowledge and recklessness may be more categorical than the Code's probabilistic definitions imply.

Data availability

All data, experimental materials, analysis code, and vignette stimuli are publicly available at `https://dgk-law-and-cognition-lab.github.io/Modeling-Guilty-Minds/`. The study was preregistered on the Open Science Framework at `https://osf.io/72eg9`.

CRediT authorship contribution statement

David G. Kamper: Conceptualization, Methodology, Formal analysis, Data curation, Visualization, Writing – original draft, Writing – review & editing. **Andrew Bistras:** Conceptualization, Methodology, Software, Investigation, Writing – original draft. **Jaime Castrellon:** Conceptualization, Methodology, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We want to thank the Castrellon Lab and the Reasoning Lab at UCLA for their feedback.

References

American Law Institute, 1985. Model Penal Code: Official Draft and Explanatory Notes. American Law Institute.

Breiman, L., 2001. Random forests. Machine Learning 45, 5–32. doi:10.1023/A:1010933404324.

Castrellon, J.J., Hakimi, S., Parelman, J.M., Lin, L., Law, J.R., Skene, J.A.G., Ball, D.A., Malekpour, A., Beskind, D.H., Vidmar, N., Pearson, J.M., Carter, R.M., Skene, J.H.P., 2022. Neural support for contributions of utility and narrative processing of evidence in juror decision making. Journal of Neuroscience 42, 7624–7633. doi:10.1523/JNEUROSCI.2434-21.2022.

Castrellon, J.J., Hakimi, S., Parelman, J.M., Lin, L., Law, J.R., Skene, J.A.G., Ball, D.A., Malekpour, A., Beskind, D.H., Vidmar, N., Pearson, J.M., Skene, J.H.P., Carter, R.M., 2023. Social cognitive processes explain bias in juror decisions. Social Cognitive and Affective Neuroscience 18, nsac057. doi:10.1093/scan/nsac057.

Christensen, R.H.B., 2019. ordinal: Regression Models for Ordinal Data. URL: https://CRAN.R-project.org/package=ordinal. r package version 2019.12-10.

Girgis, S., 2013. The mens rea of accomplice liability: Supporting intentions. Yale Law Journal 123, 460–494.

Global-Tech Appliances, Inc. v. SEB S.A., 2011. 563 U.S. 754. United States Supreme Court.

Grootendorst, M., 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794 doi:10.48550/arXiv.2203.05794.

Henrich, J., Heine, S.J., Norenzayan, A., 2010. The weirdest people in the world? Behavioral and Brain Sciences 33, 61–83. doi:10.1017/S0140525X0999152X.

Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems, pp. 4765–4774.

McInnes, L., Healy, J., Melville, J., 2018. UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 doi:10.48550/arXiv.1802.03426.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830.

Reimers, N., Gurevych, I., 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3982–3992. doi:10.18653/v1/D19-1410.

Saxe, R., Kanwisher, N., 2003. People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind". NeuroImage 19, 1835–1842. doi:10.1016/S1053-8119(03)00230-1.

Severance, L.J., Goodman, J., Loftus, E.F., 1992. Inferring the criminal mind: Toward a bridge between legal doctrine and psychological understanding. Journal of Criminal Justice 20, 107–120. doi:10.1016/0047-2352(92)90011-J.

Shen, F.X., Hoffman, M.B., Jones, O.D., Greene, J.D., Marois, R., 2011. Sorting guilty minds. New York University Law Review 86, 1306–1360.

Vilares, I., Wesley, M.J., Ahn, W.Y., Bonnie, R.J., Hoffman, M., Jones, O.D., Morse, S.J., Yaffe, G., Lohrenz, T., Montague, P.R., 2017. Predicting the knowledge–recklessness distinction in the human brain. Proceedings of the National Academy of Sciences 114, 3222–3227. doi:10.1073/pnas.1619385114.

Appendix A. Supplementary Material

Appendix A.1. Proportional odds diagnostic

Table A.1 reports the threshold-specific risk slider coefficients and ROC-derived AUC values that support the proportional odds violation described in the main text. The risk slider coefficient declines monotonically across thresholds, and discriminative accuracy (AUC) declines in parallel.

Table A.1: Threshold-specific risk slider coefficients and discriminative accuracy.

Threshold	b	95% CI	p	AUC
$\geq$ Negligence (B vs. N+)	0.059	[0.047, 0.072]	<.001	0.926
$\geq$ Recklessness (N− vs. R+)	0.052	[0.044, 0.060]	<.001	0.877
$\geq$ Knowledge (R− vs. K)	0.019	[0.011, 0.026]	<.001	0.859

Note. Each row reports results from a separate binary logistic mixed model with participant random intercepts. b = risk slider coefficient (log-odds per 1-point increase). AUC = area under the receiver operating characteristic curve. The coefficient at the Knowledge threshold is 3.2 times smaller than at the lowest threshold ($\geq$ Negligence), with non-overlapping 95% confidence intervals.

Appendix A.2. Order effects

Table A.2 reports the interaction between each vignette feature and trial position (1–5), testing whether any feature's influence on judgments changed across the five trials within a session. No interaction survived correction for multiple comparisons, indicating that the effects reported in the main text were stable across trial positions.

Table A.2: Order Effects for CLMM Models

Feature	b	SE	z	p	p_{Holm}
Source (biker)	0.068	0.089	0.76	.445	1.000
Package (duffel bag)	−0.066	0.090	−0.73	.465	1.000
Inquiry (source tells)	0.050	0.090	0.56	.579	1.000
Payment ($1,000)	0.185	0.090	2.05	.040	.241
Location (Mexico)	0.028	0.090	0.31	.753	1.000
Contents (fentanyl)	0.130	0.089	1.46	.145	.726

Note. p_{Holm} = Holm-corrected p-value across six interaction tests. No interaction survived correction. The interaction coefficient represents the change in the feature's log-odds effect per one-position increase in trial order.

Appendix A.3. Random Forest classification details

Table A.3 reports the full classification results for the three Random Forest models compared in the main text. The features-only model uses the six binary vignette features as predictors. The risk-only model uses participants' risk slider ratings as the sole predictor. The features-plus-slider model combines both predictor sets.

Table A.3: Random Forest classification accuracy (10-fold stratified cross-validation).

Model	Accuracy	κ	Chance (uniform)	Chance (majority)
Features only (6 predictors)	$.307 \pm .035$	.074	.250	.426
Risk only (1 predictor)	$.399 \pm .035$	.189	.250	.426
Features + slider (7 predictors)	$.473 \pm .029$	.254	.250	.426
Paired comparisons:				
Features vs. risk only: $t(9) = 5.86$, $p < .001$				
Features vs. combined: $t(9) = 11.15$, $p < .001$				
Risk only vs. combined: $t(9) = 6.82$, $p < .001$				

Note. Accuracy values are mean $\pm$ SD across 10 folds. κ = Cohen's kappa. Chance (uniform) = $1/4 = .250$. Chance (majority) = proportion of the most frequent category (Negligence, 42.6%).

Fig. A.1 shows confusion matrices for the features-only and features-plus-slider models. The features-only model shows diffuse misclassification across all categories, with Knowledge particularly poorly recovered (15.2% correct). Adding the risk slider substantially improves recovery of Blamelessness, Negligence, and Recklessness, but Knowledge classification does not improve and slightly decreases (12.4%). This pattern is consistent with the main-text finding that Knowledge is not organized along the same risk dimension as the other three categories.

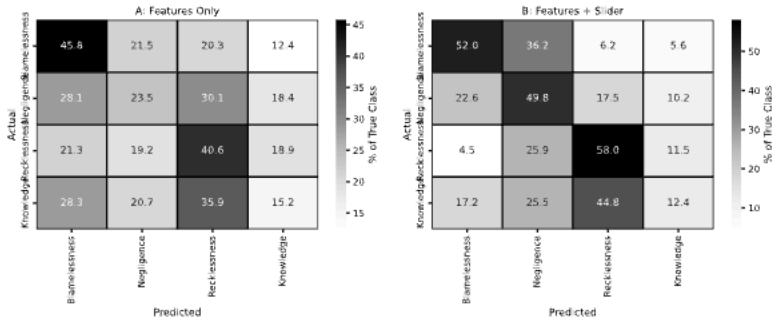


Figure A.1: Confusion matrices for the features-only (A) and features-plus-slider (B) Random Forest models. Values are percentages of each true class. Knowledge is the most difficult category to recover in both models.

Appendix A.4. SHAP feature importance

Fig. A.2 compares global SHAP importance across both models, showing the dominant role of the risk slider when included.

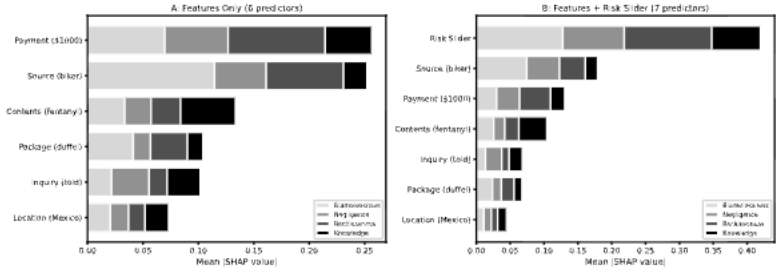


Figure A.2: Global mean |SHAP| values by feature, stacked by mental state category. (A) Features-only model. (B) Features-plus-slider model, where the risk slider dominates all vignette features.

Fig. A.3 shows the overall feature importance ranking from the features-only model, collapsed across mental state categories.

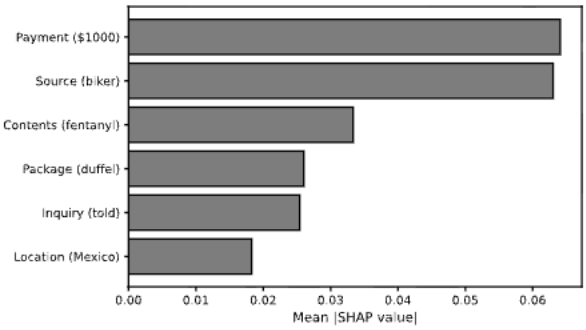


Figure A.3: Mean |SHAP| values from the features-only Random Forest, collapsed across classes. Payment and source are the two most important features overall.

Figs. A.4–A.7 show per-class SHAP beeswarm plots from the features-only model. Each point represents one trial; color indicates feature value (red = high/present, blue = low/absent). These plots complement the heatmap in the main text (Fig. 7) by showing the full distribution of SHAP values rather than only the mean.

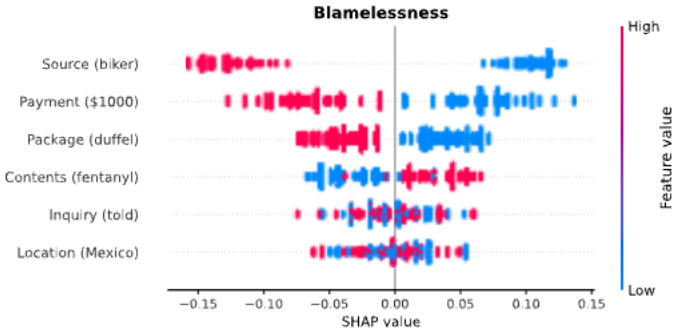


Figure A.4: SHAP beeswarm plot for Blamelessness. Source (biker) is the dominant predictor: the presence of a biker strongly pushes the model away from predicting Blamelessness (negative SHAP values for high feature value), while the uncle strongly pushes toward it.

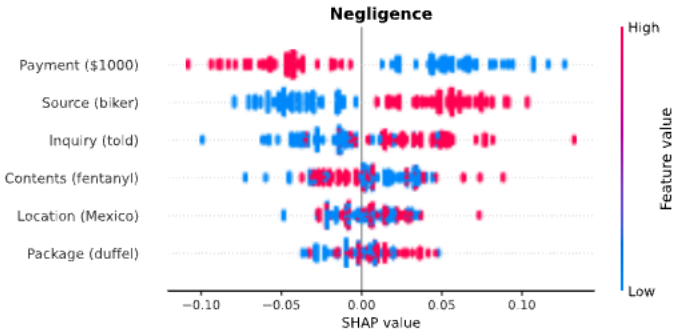


Figure A.5: SHAP beeswarm plot for Negligence. Payment is the most important feature. Source and inquiry also contribute, with more dispersed effects than for Blamelessness.

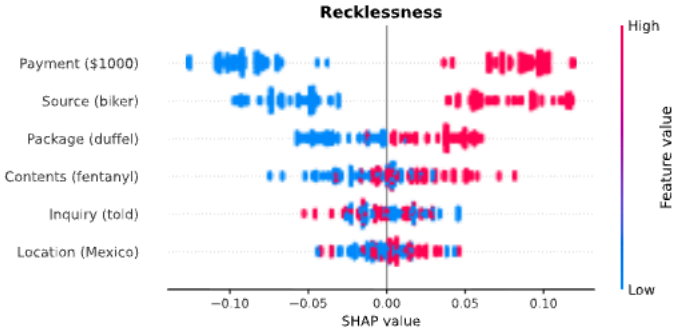


Figure A.6: SHAP beeswarm plot for Recklessness. Payment dominates, with a clear directional pattern: receiving $1,000 pushes the model toward predicting Recklessness. Source is the second most important feature.

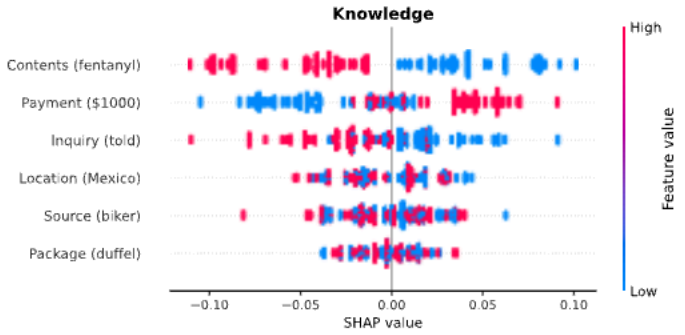


Figure A.7: SHAP beeswarm plot for Knowledge. Unlike the other three categories, contents (fentanyl) is the most important feature, consistent with the qualitative inference account discussed in the main text. Payment and inquiry also contribute, but with more dispersed and less directionally consistent effects.

Fig. A.8 shows SHAP dependence plots for the two most important features (Payment and Source) on the Recklessness class, with automatic interaction coloring.

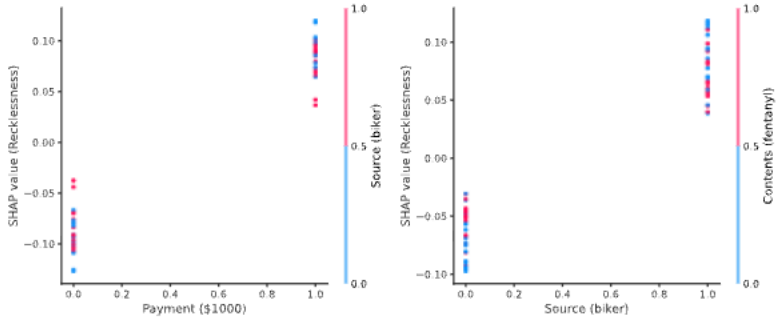


Figure A.8: SHAP dependence plots for Payment (left) and Source (right), showing their effects on the predicted probability of Recklessness. Points are colored by the strongest interacting feature.

Fig. A.9 shows the pairwise SHAP interaction matrix for the features-only model.

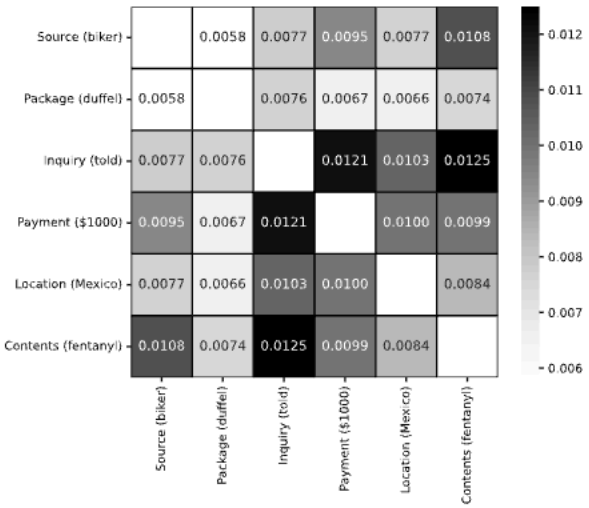


Figure A.9: Mean absolute SHAP interaction values between all pairs of vignette features. The strongest interactions involve Inquiry × Contents and Inquiry × Payment, though all interaction magnitudes are small relative to main effects.

Appendix A.5. Embedding and representational similarity details

Fig. A.10 compares per-feature embedding distances across the two sentence-transformer models. Contents and Location produce the largest embedding shifts in both models, while Payment produces one of the smallest, illustrating the dissociation between semantic prominence and behavioral importance discussed in the main text.

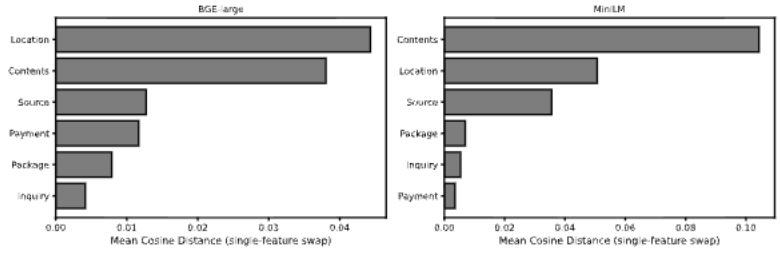


Figure A.10: Mean cosine distance between vignette embeddings that differ on a single feature, computed separately for BGE-large (left) and MiniLM (right). Features are sorted by distance magnitude within each model.

Fig. A.11 shows UMAP (McInnes et al., 2018) projections of the 64 unique vignette embeddings colored by mean culpability and by payment status.

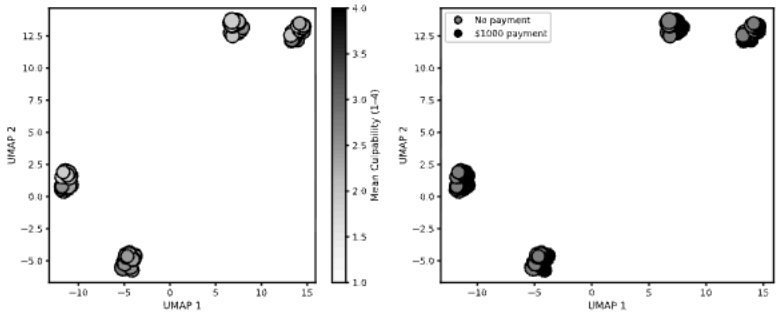


Figure A.11: UMAP projection of BGE-large vignette embeddings (64 unique configurations). Left: colored by mean culpability score (1 to 4). Right: colored by payment status. Vignettes cluster into discrete groups reflecting the embedding geometry, but within-cluster culpability variation demonstrates that semantic proximity does not determine judgment.

Fig. A.12 shows the same UMAP projection colored by each of the six binary features, revealing which features drive the embedding space geometry.

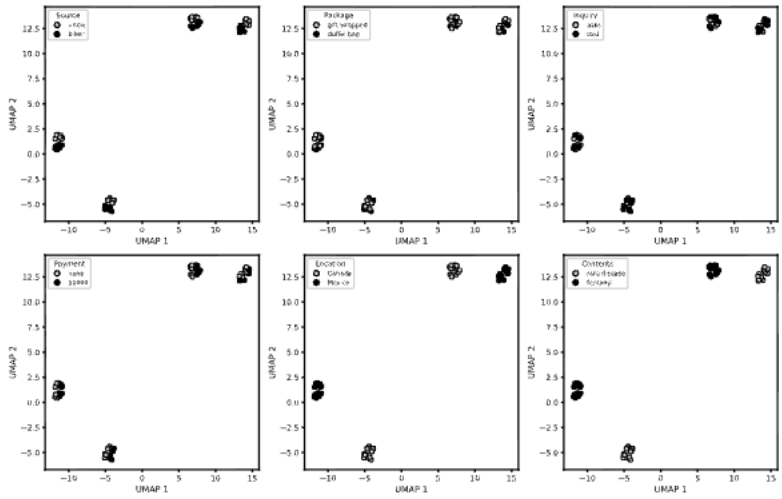


Figure A.12: UMAP vignette embeddings colored by each binary feature. Contents and Location produce the clearest cluster separation in embedding space, while Payment and Inquiry produce minimal visible separation, further illustrating the semantic-vs-behavioral dissociation.

Appendix A.6. Free-text justification analysis

Fig. A.13 shows UMAP projections of BGE-large embeddings of participants' free-text justifications, colored by mental state judgment.

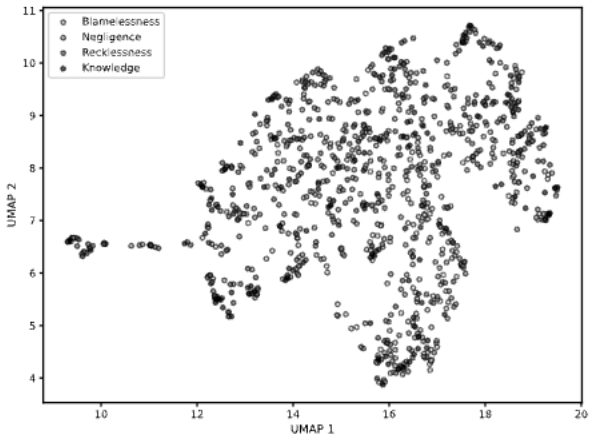


Figure A.13: UMAP projection of justification text embeddings colored by mental state judgment. The lack of clear judgment-based clustering indicates that participants used similar language to justify different mental state attributions.

Fig. A.14 shows the confusion matrix for the text-only classifier (multinomial logistic regression on BGE justification embeddings).

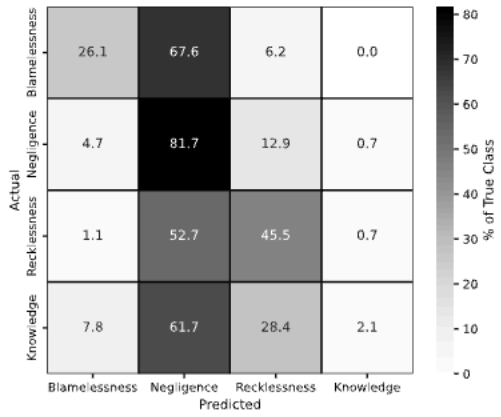


Figure A.14: Confusion matrix for the text-only classifier (BGE justification embeddings, 10-fold cross-validated). The classifier achieves 51.9% accuracy overall but almost never predicts Knowledge (2.1% recall), consistent with the finding that Knowledge attributions are driven by categorical inference that is not well captured in participants' verbal justifications.

Appendix A.7. BERTopic analysis

BERTopic analysis (Grootendorst, 2022) used MiniLM embeddings rather than BGE for topic modeling because mean-pooling architectures produce richer topic structure for short texts: the token averaging that dilutes between-vignette differences in the RSA analysis instead captures within-justification thematic content more effectively. The analysis identified 23 topics in participants' justifications. Table A.4 lists all topics with their representative keywords, dominant mental state category, and sample size.

Table A.4: BERTopic topics from free-text justifications (MiniLM embeddings, 23 topics).

Topic	Representative keywords	n	Dominant judgment	%
0	daniel, the, that, was, he	74	Negligence	42
1	money, risk, payment, asked	67	Recklessness	36
2	uncle, trusted, family, member	64	Negligence	41
3	bag, the, duffel, was, not	59	Negligence	44
4	package, he, the, was, not	55	Negligence	49
5	border, across, the, there	55	Recklessness	44
6	john, was, think, package	53	Negligence	45
7	david, uncle, would, the	46	Negligence	46
8	robert, the, that, was, of	46	Negligence	52
9	same, before, reasons, previous	34	Negligence	53
10	gift, wrapped, open, package	33	Blamelessness	45
11	andrew, was, the, that, have	29	Negligence	48
12	cannot, know, certain, knew	25	Knowledge	32
13	biker, trust, random, stranger	23	Recklessness	43
14	duffel, coffee, bag, suspicious	20	Recklessness	40
15	coffee, 1000, just, much	20	Recklessness	35
16	david, coffee, contained, bag	19	Negligence	37
17	john, coffee, transporting, bag	18	Recklessness	39
18	drugs, uncle, he, was, drug	17	Blamelessness	41
19	stranger, questions, accept, odd	17	Negligence	59
20	checked, should, have, knew	14	Negligence	50
21	coffee, uncle, believed, trusted	13	Negligence	46
22	1000, package, suspicious, dollars	12	Negligence	50

Note. Topics 0, 6, 7, 8, and 11 are partially driven by the randomized protagonist name rather than reasoning content, as indicated by proper names appearing as their top keywords. Topic 9 contains participants who wrote "same as before" or similar repeated responses. Topic 12 is the only topic where Knowledge is the dominant category; its keywords ("cannot," "know," "certain," "knew") reflect certainty-based reasoning. % = percentage of trials in the dominant category.

Fig. A.15 shows the distribution of mental state judgments across all 23 topics.

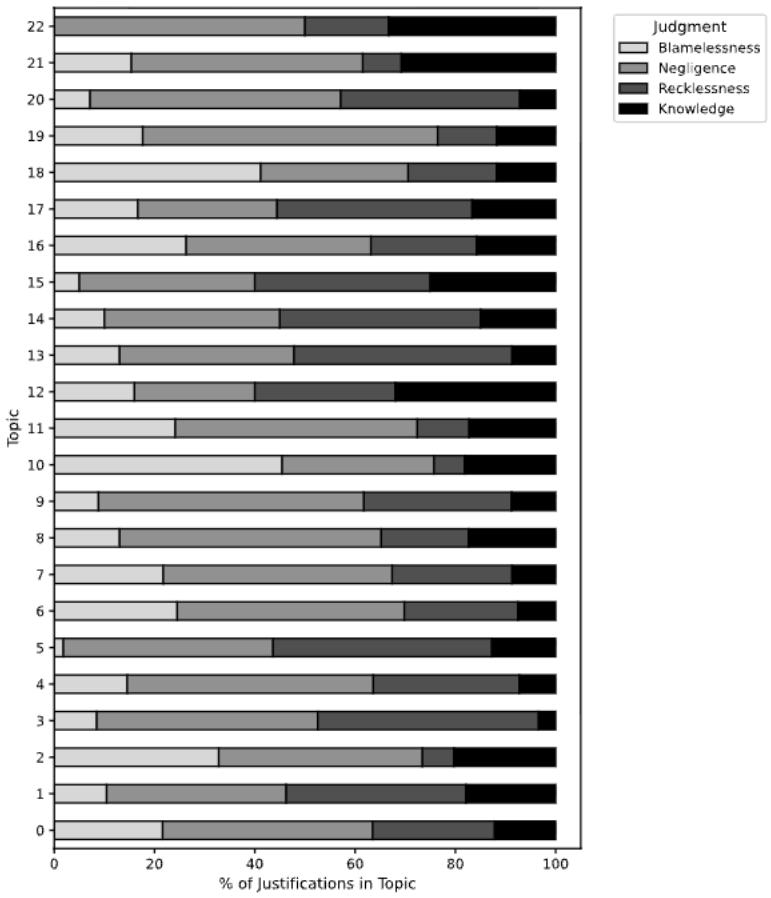


Figure A.15: Distribution of mental state judgments across BERTopic topics (MiniLM embeddings). Each bar shows the percentage of justifications in that topic assigned to each mental state category. Topics are numbered as in Table A.4. Most topics are Negligence-dominant, reflecting the overall base rate. Topics 1, 5, 13, 14, and 15 are Recklessness-dominant and reference payment, border-crossing, strangers, and suspicious packaging. Topic 12 is the only Knowledge-dominant topic.

Appendix A.8. Judgment distributions under high-awareness conditions

Table A.5 reports the distribution of mental state judgments when trials are restricted to vignettes with progressively more awareness-relevant features. As awareness cues accumulate, the modal judgment shifts from negligence to recklessness, but knowledge attributions remain at approximately 13–17% across

all conditions. This pattern supports the main-text argument that the transition from negligence to recklessness operates along a continuous risk dimension, while the transition to knowledge requires a qualitatively different inference.

Table A.5: Judgment distributions (%) under progressively stricter high-awareness vignette filters.

Filter	N	Blamelessness	Negligence	Recklessness	Knowledge
All trials	1,060	16.7	**42.6**	27.0	13.7
Payment = \$1,000	538	12.1	35.1	**36.8**	16.0
Source = biker	521	7.5	**47.2**	32.6	12.7
Payment + biker	265	4.5	38.5	**41.1**	15.8
Payment + biker + asks	128	3.9	34.4	**44.5**	17.2
All four cues	67	6.0	32.8	**44.8**	16.4

Note. Bold values indicate the modal category within each filter. "Asks" = protagonist actively asked about the package contents. "All four cues" = payment + biker + asks + fentanyl. As awareness-relevant features accumulate, the modal judgment shifts from negligence to recklessness, but knowledge remains at approximately 13–17% across all conditions.