APPENDIX: Supplementary Material for *Convergent Flexibility: How International Law Keeps Pace with Technological Change*

Justin Key Canfil *Carnegie Mellon University*

A.1 Model Diagnostics and Sensitivity **i**

A.1.1 Statistical Power i

A.1.2 Randomization Inference ii

A.1.3 Balance on Observables iii

A.1.4 LASSO Regularization iv

A.1.5 Numerical Results vi

A.2 Sample Characteristics **ix**

A.2.1 Procedure and Wave Stability xi

A.2.2 Nonparticipation xiii

A.2.3 Engagement and Manipulation Checks xiii

A.3 Heterogeneous Effects **xvi**

A.4 Theoretical Mechanisms **xx**

A.4.1 Structural Topic Models xx

A.4.2 Themes and Excerpts xxiii

A.4.3 Mediation Analysis xxvi

A.5 Codebook **xxviii**

A.5.1 Derivative Covariates xxviii

A.5.2 Secondary Outcomes xxxii

A.6 Appendix Bibliography **xxxiv**

A.1 Model Diagnostics and Sensitivity

A.1.1 Statistical Power

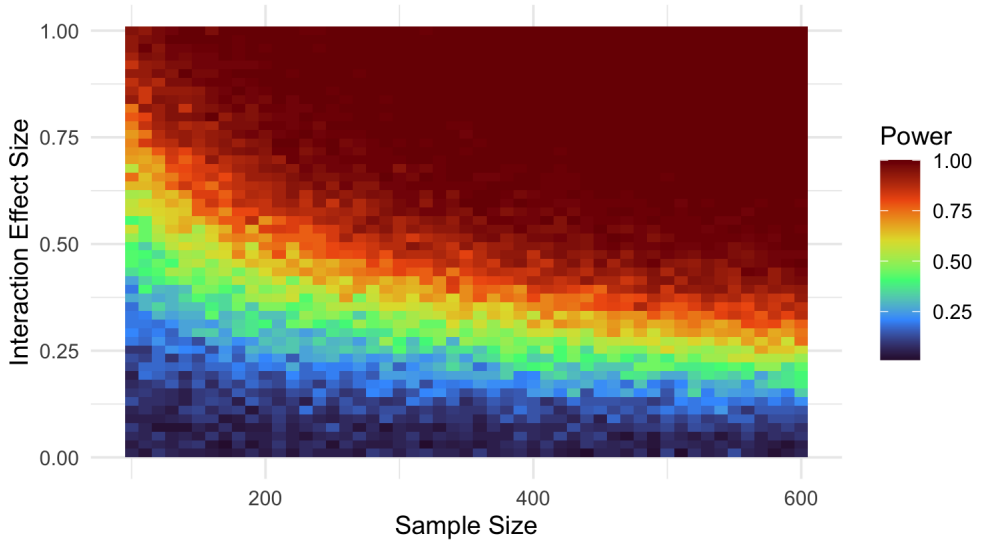
Statistical power is crucial because it measures a study's ability to detect an effect if one truly exists, reducing the risk of committing a Type II error, where a real effect is falsely deemed non-significant. A high statistical power ensures that the study can reliably identify true effects, increasing the confidence in the results and the overall validity of the research conclusions. For factorial experiments, power can decrease dramatically as the number of treatments increases [Baranger et al. 2023; Sommet et al. 2023]. While the risk of Type I errors is negligible, Gelman and Carlin [2014] and Muralidharan et al. [2023] point out the risk of Type-S (sign) and -M (magnitude) errors in low-powered environments.

Power analysis was conducted prior to preregistration to determine the target sample size. A total of $127,500$ sample/effect size permutations were analyzed with $1,000$ simulations per scenario. Figure A.1 plots the relationship between sample size, interaction effect size, and power with the effect of lower-order coefficients averaged. A threshold of $p = 0.05$ is used. A priori, the preanalysis plan assumed a conservative effect size ($\beta_{T1\times T2} < 0.5$), pinpointing a target sample size of $N = 400$.

The observed effect size—$\beta_{T1\times T2} = |5.191|$ to $|5.256|$—is far larger than the minimum effect size required to recover a significant effect for $N = 400$ participants, as shown in Figure A.1 ($\beta_{T1\times T2} < 1$). We should be further reassured of adequate statistical power for the following reasons:

- Participants consisted of highly educated elites who were highly motivated to take part in a professionally familiar simulation. (Very high pass rate on attention checks.) Thus, a very large observed effect sizes (and high R^2) is not unexpected.
- The target sample size was exceeded ($N > 400$).
- A significant effect was recovered for the hypothesized interaction $T_1 \times T_2$.

Figure A.1: Heatmap of Factorial Statistical Power



Note: Factorial power analysis based on 127,500 scenarios and 1,000 simulations per scenario. Main interaction effect only. Lower-order coefficients are averaged for plotting purposes. The supplementary material contains additional code to generate plots with variable main effects.

- The signs for T_2 and $T_1 \times T_2$ are in the predicted direction.

Moreover, a high degree of contextual complexity has even been found to dampen observed effect sizes [Brutger and Kertzer 2018]. Since the current experiment is extremely complex, this implies that, if anything, the true effect may be even larger. The larger the true effect size, the lower the requisite sample size. Holding constant sample size at $N = 400$, power is increased. Thus, we should conclude that the study was adequately powered.

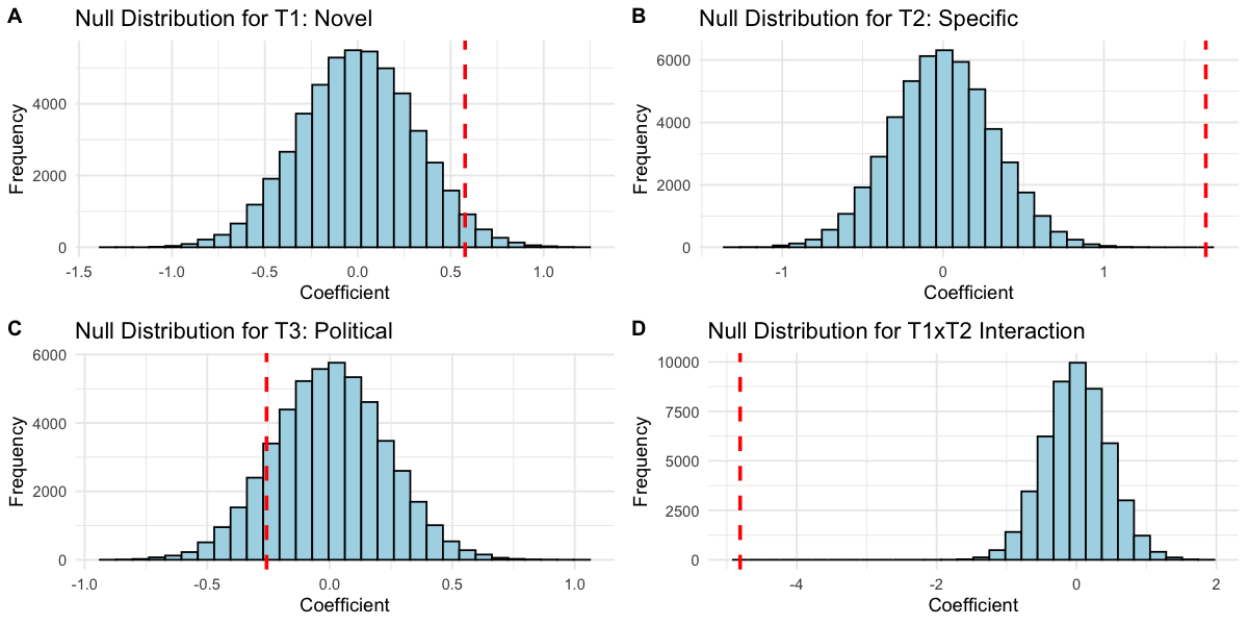
A.1.2 Randomization Inference

Randomization inference is employed to test whether the observed effect might be due to random variation. Randomization inference involves simulating the assignment process multiple times, keeping the treatments constant but randomly reassigning which units receive each treatment to generate a distribution of possible outcomes that could have occurred by chance. By comparing the observed treatment effect to this distribution, it is possible to

determine whether the effects seen in the experiment are statistically significant or likely due to random chance. The proportion of simulated scenarios where the effect is as extreme as, or more extreme than, the observed effect is used to calculate a p-value, indicating the likelihood of observing such an effect if there were truly no effect of the treatment.

To assess the validity of the estimates, I conduct $50,000$ re-randomization permutations for each treatment factor (T_1, T_2, and T_3), plus the preregistered interaction between T_1 and T_2. The null distribution is extracted from the results. I then calculate the p-values for each effect. Results are plotted in Figure A.2. The results confirm that the coefficients on $T_1 + T_2 + (T_1 \times T_2)$ are significantly different from mere chance coefficients. The interaction coefficient is -4.81 ($p < 0.001$). As with the OLS regressions, we confirm T_3 does not depart from the null ($p = 0.255$).

Figure A.2: Factorial Randomization Inference



A.1.3 Balance on Observables

Under random assignment, participants in all $2^3 = 8$ experimental conditions are identically distributed in expectation. However, because the study used simple random assignment,

imbalance is possible in smaller sample sizes, hence why the study included multiple waves. I use love plots, shown in Figure A.3, to detect imbalances on observables across treatment conditions. The participant pool exhibits generally good covariate balance.

There is some evidence that assignment differences were significant for two variables in the ancillary treatment, T_3 (client type). These include news consumption frequency (from daily to never) and news intensity (a derivative measure developed by multiplying news consumption frequency by whether consumption includes relevant news topics, such as science and technology).

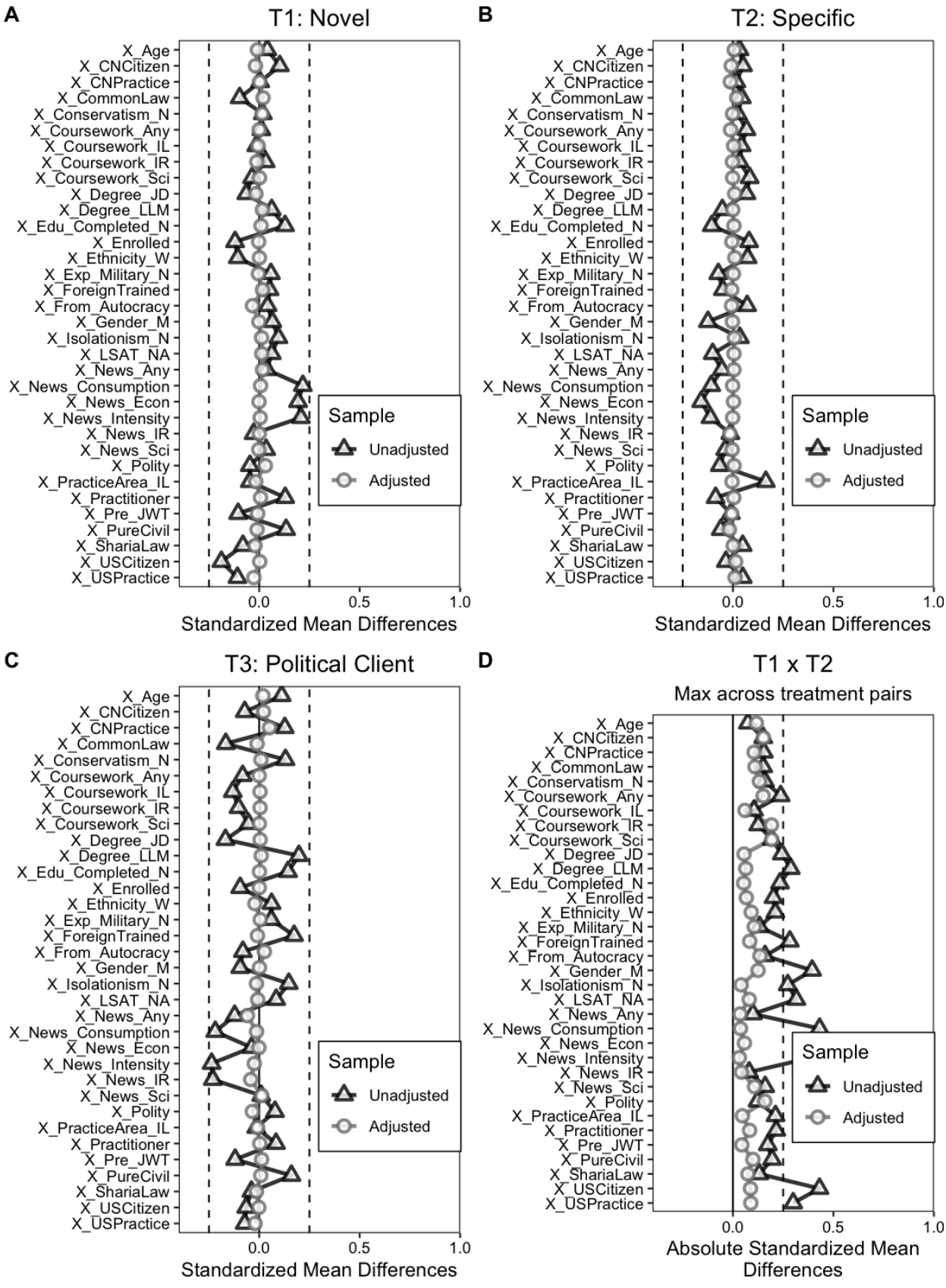
This could be a problem for the research if attention to the news moderates the outcome—for example if a real satellite controversy became salient for participants while the experiment was in progress. At least two satellite breakthroughs did occur between 2020 and 2024, but neither occurred while an experimental wave was underway. These include China's test of a Fractional Orbital Bombardment (FOBS) hybrid system in November 2021; and the launch of the James Webb Telescope, which uses a similar Lissajous orbit, in December 2021. Both events captured significant public attention. I include a dummy variable (`X_Pre_JWT`) for comparing outcomes before and after exposure to these events.

Covariate adjustment is sufficient to resolve imbalances, including joint imbalances across all treatment pairs. As Table A.1 shows, the main results are robust to a variety of adjustment specifications.

A.1.4 LASSO Regularization

LASSO regularization is a statistical method used in regression models to enhance prediction accuracy and interpretability. LASSO helps in selecting only the most relevant covariates by shrinking less important coefficients to zero, effectively reducing overfitting and improving model performance. This method simplifies the model by excluding non-informative covariates, making it particularly useful in experiments with a large number of potential

Figure A.3: Covariate Balance (Love Plots)



predictors. LASSO has a no-missingness requirement. Consistent with best practices [Montgomery et al. 2018], this is enforced by pretreatment covariate measurement.[22] LASSO also requires the researcher to choose a regularization parameter, λ, that determines the severity of the inclusion penalty. Results are reported in Figure A.4.

First, the sample is randomly divided into training/test sets (90%/10%). Second, an elastic net approach is used to confirm whether pure LASSO ($\alpha = 1$) is optimal. Consistent with the preregistered preanalysis plan, I choose the λ that minimizes the 10-fold cross-validation error averaged over 10 runs. Next, the results are used to generate out-of-sample predictions on the test set. The MSE for the training and test sets are comparable. Finally, I use a bootstrapping approach ($50,000$ iterations) to estimate the distribution of differences between the test set MSE and cross-validated error. Figure A.4 plots this distribution. Each bar represents the frequency of differences in a specified range. The confidence intervals (95%) capture the range in which the true mean differences is located. If the confidence interval includes zero, the model is well-fitted.

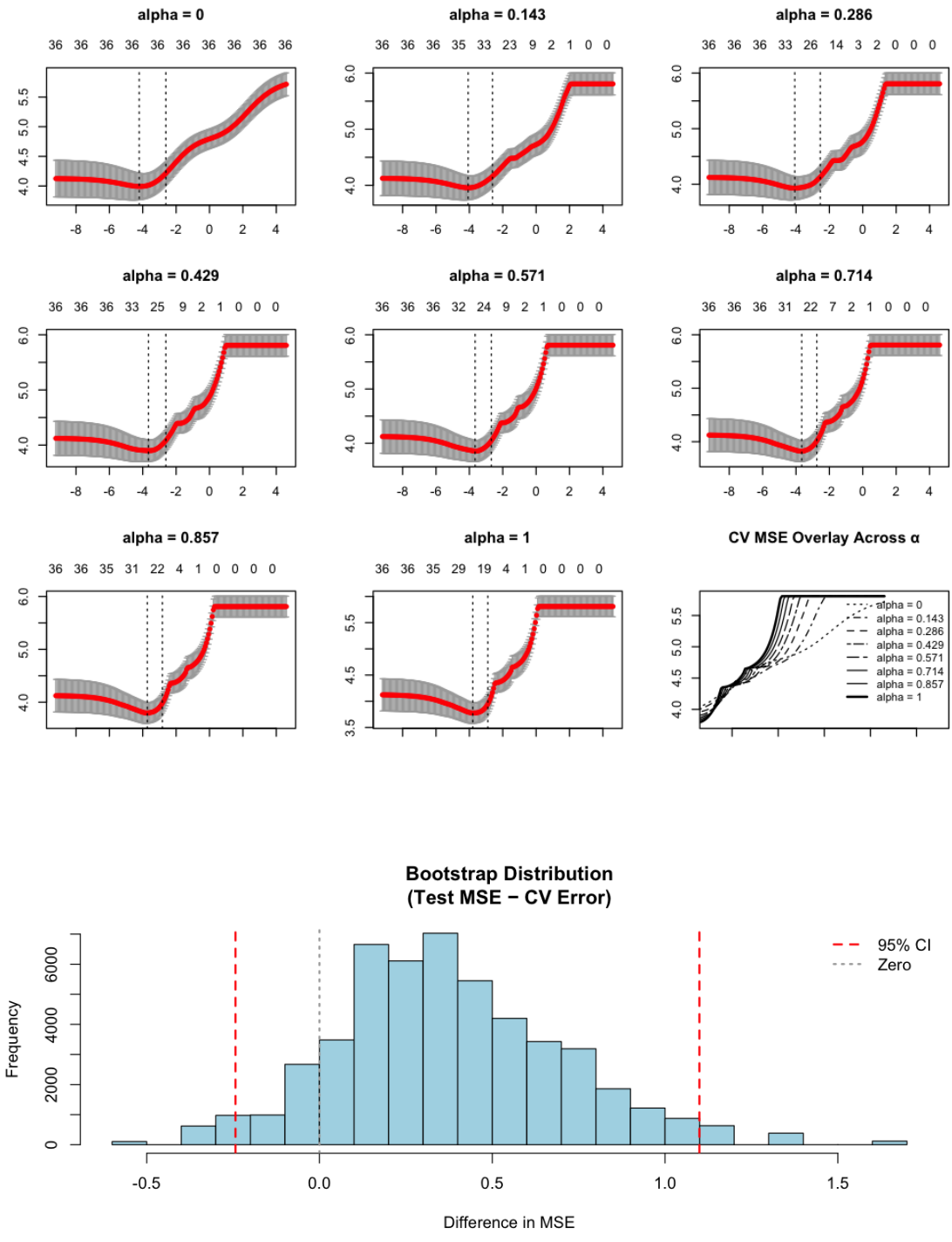
For transparency, I report both the LASSO and base results in Table A.1. The main results are robust to model specification.

A.1.5 Numerical Results

This subsection contains numerical results for the estimated ATE in Figure 6. Additional models with alternative specifications are also included for robustness. A.5 describes several secondary outcomes based on response metadata.

[22] `LSAT Score` and `Aversion` questions are excluded from the procedure. Foreign-trained lawyers (eg. LLM degree-holders) may not have taken the LSAT exam, in which case their LSAT score is NA. Aversion covariates were also introduced late. As aversion measures are available for only 20% of the sample and were not preregistered, they are not included.

Figure A.4: LASSO cross-validation and bootstrapping test

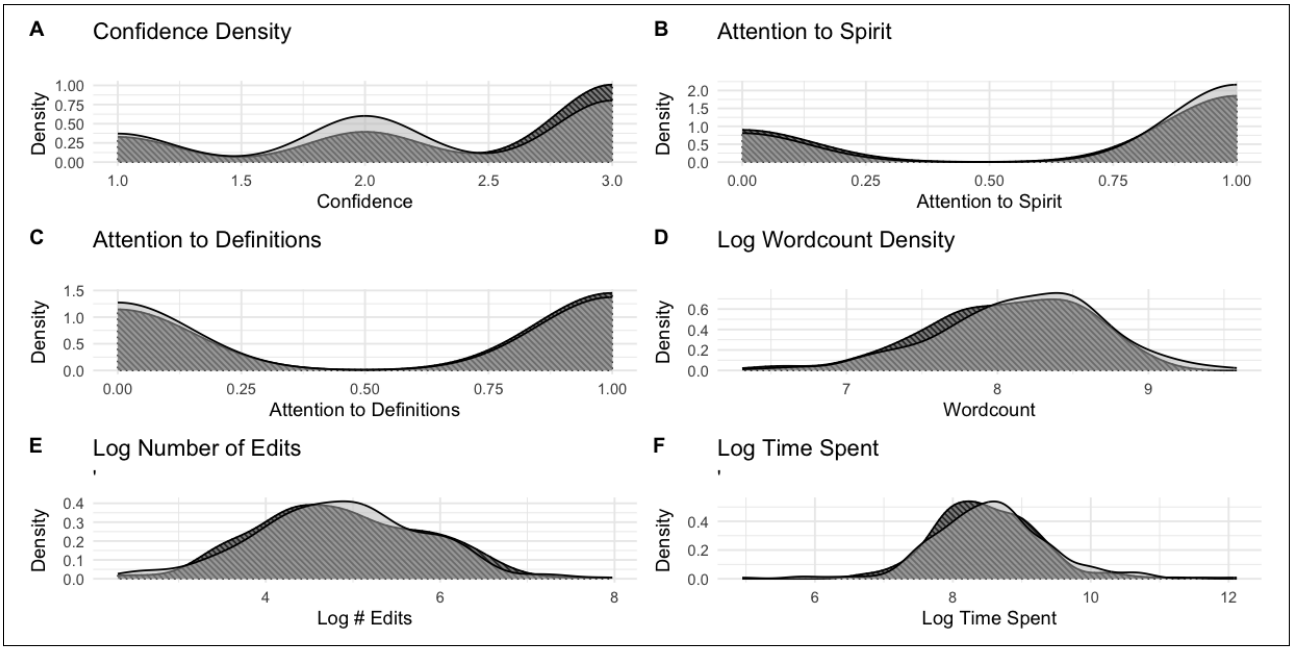


Note: Above: Elastic net for α and λ optimization. Y-axis is Mean Squared Error (MSE) and X-axis is λ value. Below: Bootstrapped distribution of differences to test for under/overfitting based on optimal α, λ.

Table A.1: Summary Statistics for Four OLS Models

	Perceived Restrictiveness of Article II			
	OLS-Unadjusted	OLS 3W Interaction	LASSO-Adjusted	General Only
	(1)	(2)	(3)	(4)
Constant	0.559** (0.192)	0.746** (0.244)	1.424 (1.691)	0.290 (0.657)
T_Specific	1.634*** (0.248)	1.310*** (0.353)	1.646*** (0.253)	
T_Newtech	0.577* (0.242)	0.327 (0.351)	0.535* (0.245)	0.584* (0.293)
T_Political	−0.258 (0.176)	−0.619+ (0.339)	−0.272 (0.180)	0.170 (0.848)
X_Age			0.023 (0.025)	
X_CNCitizen			0.463 (0.503)	
X_CNPractice			0.554 (0.789)	
X_CommonLaw			−0.271 (1.143)	
X_From_Autocracy				0.703 (0.765)
X_Conservatism_N			−0.088 (0.071)	−0.084 (0.182)
X_Coursework_IL			−0.078 (0.119)	
X_Coursework_IR			−0.092 (0.133)	
X_Coursework_Sci			0.105 (0.156)	
X_Degree_JD			−0.614+ (0.331)	
X_Gender_M			0.230 (0.196)	
X_Isolationism_N			0.071 (0.057)	0.235+ (0.139)
X_News_Any			−0.841 (0.960)	
X_News_Econ			0.296 (0.195)	
X_News_Intensity			0.006 (0.112)	
X_News_IR			0.548 (0.445)	
X_PracticeArea_IL			0.027 (0.188)	
X_Practitioner			−0.200 (0.228)	−0.370 (0.435)
X_Pre_JWT			−0.485* (0.193)	
X_PureCivil			−0.779 (1.184)	
X_ShariaLaw			0.271 (0.785)	
T_Specific:T_Newtech	−4.810*** (0.353)	−4.423*** (0.508)	−4.814*** (0.362)	
T_Specific:T_Political		0.639 (0.496)		
T_Newtech:T_Political		0.482 (0.485)		
T_Specific:T_Newtech:T_Political		−0.758 (0.708)		
T_Political:X_From_Autocracy				−0.913 (1.089)
T_Political:X_Conservatism_N				0.131 (0.234)
T_Political:X_Isolationism_N				−0.291 (0.185)
T_Political:X_Practitioner				−0.006 (0.607)
Observations	450	450	450	239
R^2	0.399	0.402	0.432	0.045
Adjusted R^2	0.394	0.392	0.399	0.003
Residual Std. Error	1.870 (df = 445)	1.872 (df = 442)	1.861 (df = 425)	2.237 (df = 228)
F Statistic	73.970*** (df = 4; 445)	42.420*** (df = 7; 442)	13.450*** (df = 24; 425)	1.082 (df = 10; 228)
Note:				+ p<0.1; * p<0.05; ** p<0.01; *** p<0.001

Figure A.5: Change in Secondary Outcomes as T_2: Specificity Varies



Note: Disaggregated by T_2 (specificity) condition. Solid ($T_2 = 0$, general) vs. hatched ($T_2 = 1$, specific).

A.2 Sample Characteristics

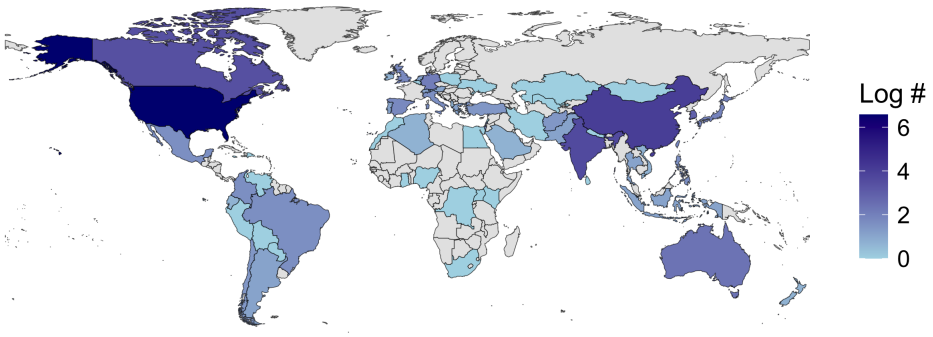
Sixty-two countries are represented in the sample. Log counts per country are plotted in Figure A.6. Using data from Nyrup et al. [2025] on the educational background of cabinet-level officials from 141 countries, I derive an indicator of a country's "lawyerliness" by calculating the proportion of all officials in a given country who had terminal law degrees versus other degrees. I then regress `X_LawPtP` on a dummy variable `in_sample` indicating whether that country was included in the current sample. On a range of zero to 0.911, sampled countries score 0.174 higher in "lawyerliness" than non-sampled countries on average, and the difference is significant ($p < 0.001$). However, the baseline across all countries is already 0.474 (also $p < 0.001$), implying that levels of "lawyerliness" are not substantially lower in the rest of the world.

Table A.2: Descriptive Information About Treated Participants

Variable	N	Median	Std Dev	Min	Max	Levels
Age	450	25.88	25.00	3.89	19.00	54.00
CNCitizen	450	0.06	0.00	0.24	0.00	1.00
CNPractice	450	0.02	0.00	0.13	0.00	1.00
CommonLaw	450	0.82	1.00	0.39	0.00	1.00
Conservatism_N	450	2.68	2.00	1.31	1.00	7.00
Coursework_Any	450	0.94	1.00	0.23	0.00	1.00
Coursework_IL	450	1.42	2.00	0.81	0.00	2.00
Coursework_IR	450	1.30	1.00	0.75	0.00	2.00
Coursework_Sci	450	0.29	0.00	0.59	0.00	3.00
Degree_JD	450	0.84	1.00	0.37	0.00	1.00
Degree_LLM	450	0.19	0.00	0.39	0.00	1.00
Edu_Completed_N	450	2.28	2.00	1.15	1.00	4.00
Enrolled	450	0.60	1.00	0.49	0.00	1.00
Ethnicity_W	450	0.51	1.00	0.50	0.00	1.00
Exp_Military_N	450	0.09	0.00	0.28	0.00	1.00
ForeignTrained	450	0.17	0.00	0.38	0.00	1.00
From_Autocracy	450	0.09	0.00	0.29	0.00	1.00
Gender_M	450	0.54	1.00	0.50	0.00	1.00
Isolationism_N	450	2.88	3.00	1.62	1.00	7.00
LawPtP	449	0.80	0.88	0.15	0.25	0.91
LSAT_NA	450	0.19	0.00	0.39	0.00	1.00
LSAT_Score	366	169.04	171.00	9.26	120.00	180.00
News_Any	450	0.99	1.00	0.10	0.00	1.00
News_Consumption	450	3.03	3.00	0.95	0.00	4.00
News_Econ	450	0.44	0.00	0.50	0.00	1.00
News_Intensity	450	3.02	3.00	0.97	0.00	4.00
News_IR	450	0.94	1.00	0.24	0.00	1.00
News_Sci	450	0.49	0.00	0.50	0.00	1.00
Polity	450	6.91	8.00	4.19	-10.00	10.00
PracticeArea_IL	450	0.47	0.00	0.50	0.00	1.00
Practitioner	450	0.41	0.00	0.49	0.00	1.00
Pre_JWT	450	0.53	1.00	0.50	0.00	1.00
PureCivil	450	0.17	0.00	0.38	0.00	1.00
ShariaLaw	450	0.02	0.00	0.15	0.00	1.00
USCitizen	450	0.69	1.00	0.46	0.00	1.00
USPractice	450	0.88	1.00	0.33	0.00	1.00
Treated	450	1.00	1.00	0.00	1.00	1.00
Wave_Fall 2020	450	0.42	0.00	0.49	0.00	1.00
Wave_Fall 2024	450	0.32	0.00	0.47	0.00	1.00
Wave_Pilot 2020	450	0.11	0.00	0.31	0.00	1.00
Wave_Spring 2024	450	0.07	0.00	0.25	0.00	1.00
Wave_Spring 2025	450	0.08	0.00	0.27	0.00	1.00

Note: Aversion measures were added to later waves. Discrepancies in N for `LSAT_NA` are due to the fact that not all law programs require the LSAT. In addition, one participant was from a country not included in the `X_LawPtP` data (Mongolia).

Figure A.6: Geographic composition of the sample (home/practice countries)



Note: Counts are logged to account for skewness in the sample.

A.2.1 Procedure and Wave Stability

Before taking part, participants were required to register for the contest by completing a 15-question pretreatment survey including demographic, ideological, career, and knowledge questions.[23] After completing the survey, a Qualtrics link containing case materials was emailed to the law school `.edu` email address with which they registered. As a security measure, participants were required to authenticate with the same `.edu` email in order to access the case materials. To ensure written responses were authentic, measures were implemented to prevent the use of generative AI such as ChatGPT.[24]

Consent was obtained during registration. Participants were told that only that the case would involve a scenario in international relations. After completing the pretreatment survey, prospective participants received an email containing a link to the contest materials. Participants authenticated by confirming the `.edu` email address they originally registered with. Participants were instructed to proceed only when they would have at least 30

[23] In later waves, additional questions intended to gauge technology aversion, risk aversion, and ambiguity aversion are added to the survey. Data on these questions are available for 20 percent of the treated sample.

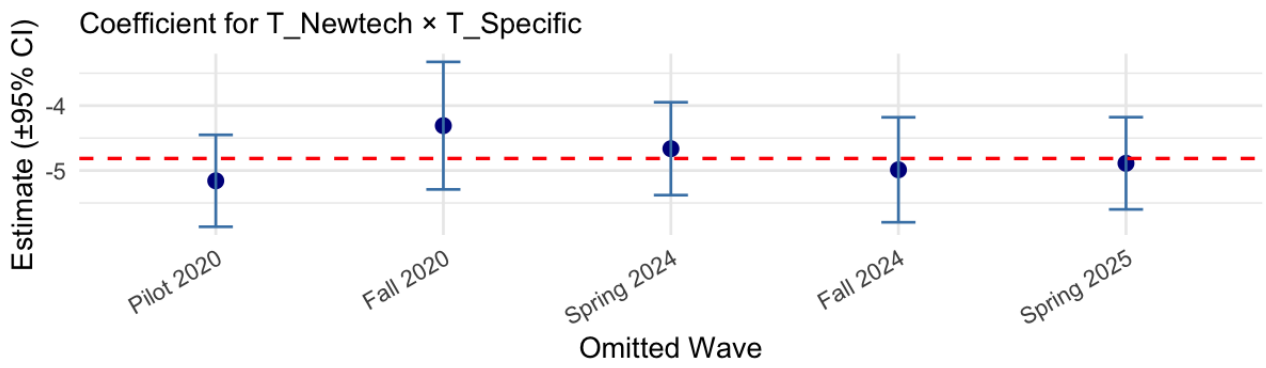
[24] The copy/paste feature was disabled at the browser level via Javascript. Prompts and case information was also embedded as an image to hinder text copying. The right-click-save feature was also disabled to prevent users from saving the prompts as an image.

uninterrupted minutes to complete the exercise and to use a computer (not phone). Participants were also told that outside research would not be necessary because the relevant law was hypothetical and that assumptions based on external information would be considered disqualifying. Participants who agreed were asked to confirm their email address a third time before proceeding. After clicking through background materials, subjects who agreed to view the case facts were randomly assigned to one of eight treatment conditions.

A preregistered pilot study ($N = 48$) was also conducted in August 2020 via the closed Reddit communities `r/lawyers` (99k members) and `r/lawschool` (794k members). Each subreddits receive upwards of 180,000 unique hits per month. Law students and lawyers are required to verify themselves with a valid US law school email address and/or State Bar ID. Subreddit moderators conduct verification independently and do not share this information. Therefore, as a secondary security measure, pilot participants were required to (a) register for the study using a law school `.edu` email address and (b) upload a law school transcript before proceeding. All participants were able to verify their eligibility. The data have been anonymized to protect subject/institutional identities. Pilot responses are included in the analysis by default.

I assessed whether the main results were stable across waves using a leave-one-out (jackknife) robustness test and a set of formal statistical checks. The jackknife plot in Figure A.7 shows that the estimated interaction coefficient remains large, negative, and nearly identical regardless of which semester is excluded; all 95% confidence intervals overlap substantially, indicating little sensitivity to temporal composition. The jackknife variance ratio ($0.70 < 1$) confirms that observed variation across semesters is smaller than expected from sampling noise. Taken together, the evidence demonstrates that the negative interaction effect of `T_Newtech × T_Specific` is statistically robust across time, including for the pilot, with no single wave exerting undue influence on the overall result.

Figure A.7: Leave-one-out (jackknife) robustness check



Note: Sign and magnitude of main result coefficients are statistically indistinguishable between waves.

A.2.2 Nonparticipation

Table A.3 lists summary statistics for untreated participants only. A total of $N = 1,147$ individuals registered their interest in the contest by completing the 15-question pretreatment questionnaire. Participants in Table A.3 completed the questionnaire but did not ultimately enroll in the contest. Although there was no attrition, this information is included for greater transparency about the sampling procedure.

A.2.3 Engagement and Manipulation Checks

Cognitive burden was a primary concern during the experimental design phase. However, I was concerned that omitting detail would threaten the validity of the results, either because lawyers are trained to probe for textual weaknesses, or because a simpler version would reduce realism. Facing this tradeoff, I opted for greater complexity, and to reduce inattention and attrition, I relied on a contest scheme with large prizes.

Four manipulation checks were performed:

1. Ability to recall and correctly characterize the shape of the satellite's orbital path;

Table A.3: Descriptive Information About Known Nonrespondents

Variable	N	Median	Std Dev	Min	Max	Levels
Age	586	26.57	26.00	6.35	0.00	99.00
CNCitizen	697	0.04	0.00	0.21	0.00	1.00
CNPractice	585	0.01	0.00	0.10	0.00	1.00
CommonLaw	585	0.80	1.00	0.40	0.00	1.00
Conservatism_N	565	2.64	2.00	1.40	1.00	7.00
Coursework_Any	697	0.80	1.00	0.40	0.00	1.00
Coursework_IL	697	1.28	2.00	0.90	0.00	2.00
Coursework_IR	697	1.05	1.00	0.86	0.00	2.00
Coursework_Sci	697	0.25	0.00	0.57	0.00	3.00
Degree_JD	664	0.72	1.00	0.45	0.00	1.00
Degree_LLM	664	0.15	0.00	0.36	0.00	1.00
Edu_Completed_N	575	2.55	3.00	1.12	1.00	4.00
Enrolled	575	0.49	0.00	0.50	0.00	1.00
Ethnicity_W	586	0.49	0.00	0.50	0.00	1.00
Exp_Military_N	660	0.08	0.00	0.26	0.00	1.00
ForeignTrained	697	0.13	0.00	0.34	0.00	1.00
From_Autocracy	585	0.08	0.00	0.27	0.00	1.00
Gender_M	586	0.49	0.00	0.50	0.00	1.00
Isolationism_N	565	2.87	3.00	1.60	1.00	7.00
LawPtP	585	0.81	0.88	0.15	0.19	0.90
LSAT_NA	697	0.33	0.00	0.47	0.00	1.00
LSAT_Score	468	167.77	171.00	11.00	120.00	180.00
News_Any	697	0.80	1.00	0.40	0.00	1.00
News_Consumption	565	3.08	3.00	0.92	0.00	4.00
News_Econ	697	0.39	0.00	0.49	0.00	1.00
News_Intensity	565	3.06	3.00	0.96	0.00	4.00
News_IR	697	0.76	1.00	0.43	0.00	1.00
News_Sci	697	0.41	0.00	0.49	0.00	1.00
Polity	585	7.13	8.00	3.97	-10.00	10.00
PracticeArea_IL	697	0.34	0.00	0.47	0.00	1.00
Practitioner	575	0.52	1.00	0.50	0.00	1.00
Pre_JWT	697	0.47	0.00	0.50	0.00	1.00
PureCivil	585	0.18	0.00	0.38	0.00	1.00
ShariaLaw	585	0.02	0.00	0.15	0.00	1.00

Note: These individuals began a separate registration survey but ultimately did not participate in the experiment. Discrepancies in the nonparticipant set size (N) are a function of dropouts at various points during the pretreatment questionnaire.

2. Ability to recall the treaty's upper limit on the number of "permitted" satellites each country may have;

3. Explicit reference to the operative provision of the treaty (Article II). (Note: because reference to Article II is technically subjective, this is probably not a reliable manipulation check.)

4. Signaled potential condition awareness by mentioning "ambiguity."

Pass rates were as follows: 89.1% on the first check, 99.3% on the second, 89.8% on the third, and 94.4% on the fourth. No subject failed all three checks. Engagement was also measured in several ways:

- **Participants recruitment and aptitude:** Participants were recruited based on their affiliation with highly selective law schools. Participants displayed a high degree of cognitive aptitude, as reflected in their LSAT scores (mean score = 171).

- **Engagement with case facts:** Written briefs demonstrate deep engagement with the case, including detailed comparisons between the technology treatment (T1) and past technology examples in Schedule I. Briefs also show careful evaluation of how treaty text (T2) maps onto T1, often quoting relevant passages verbatim.

- **Selective attention:** Participants selectively ignored extralegal "red herrings," such as the satellite's reported distance from Earth, which was highlighted in the visualization but had no actual bearing on interpretation under Article II of the treaty.

- **Behavioral measures:** Figure 14 in the appendix plots the density of secondary outcomes, including reported confidence in interpretation; time spent reading the prompt; time spent writing the brief; brief wordcount; attention to definitions; attention to extralegal factors (e.g., citations to the preamble); and the number of edits made during writing (proxied by the number of clicks in the long-form text box). Confidence was slightly higher in the specific condition ($T2 = 1$), while wordcount

and time spent were slightly lower. This was expected, since specificity was theorized to make the law's application clearer. No substantial differences were observed across other measures.

- **Mediation analysis:** Section D.3 of the appendix reports mediation analyses testing whether variation in attention to treaty definitions, importance assigned to specific provisions, emphasis on extralegal factors, or written response length mediated the results. None of these effects was significant.

- **AI use controls:** Use of generative AI is not suspected. The first three waves of the experiment were fielded in 2020, prior to the launch of ChatGPT (GPT-3) in November 2021. Subsequent waves employed JavaScript to prevent right-clicking and copy-paste operations in the web browser.

An unrelated concern was that participants in later waves might be familiar with the technology in question (a Lissajous orbit), potentially biasing results. This concern arose because the widely publicized James Webb Space Telescope—launched in December 2021—reportedly uses a Lissajous orbit. However, the median participant reported consuming no science news, and a dummy variable for pre- and post-telescope launch (Appendix, p. v) indicates no significant effect on interpretation.

A.3 Heterogeneous Effects

To measure subgroup heterogeneity, I employ Bayesian Additive Regression Trees (BART). BART is a machine learning approach recommended for automating the discovery of "important" predictor variables. BART is recommended for the automated discovery of heterogeneous treatment effects in experimental contexts. As Green and Kern [2012] explain, one of BART's key advantages is that it "automates the detection of nonlinear relationships and interactions, thereby reducing researchers' discretion when analyzing experimental

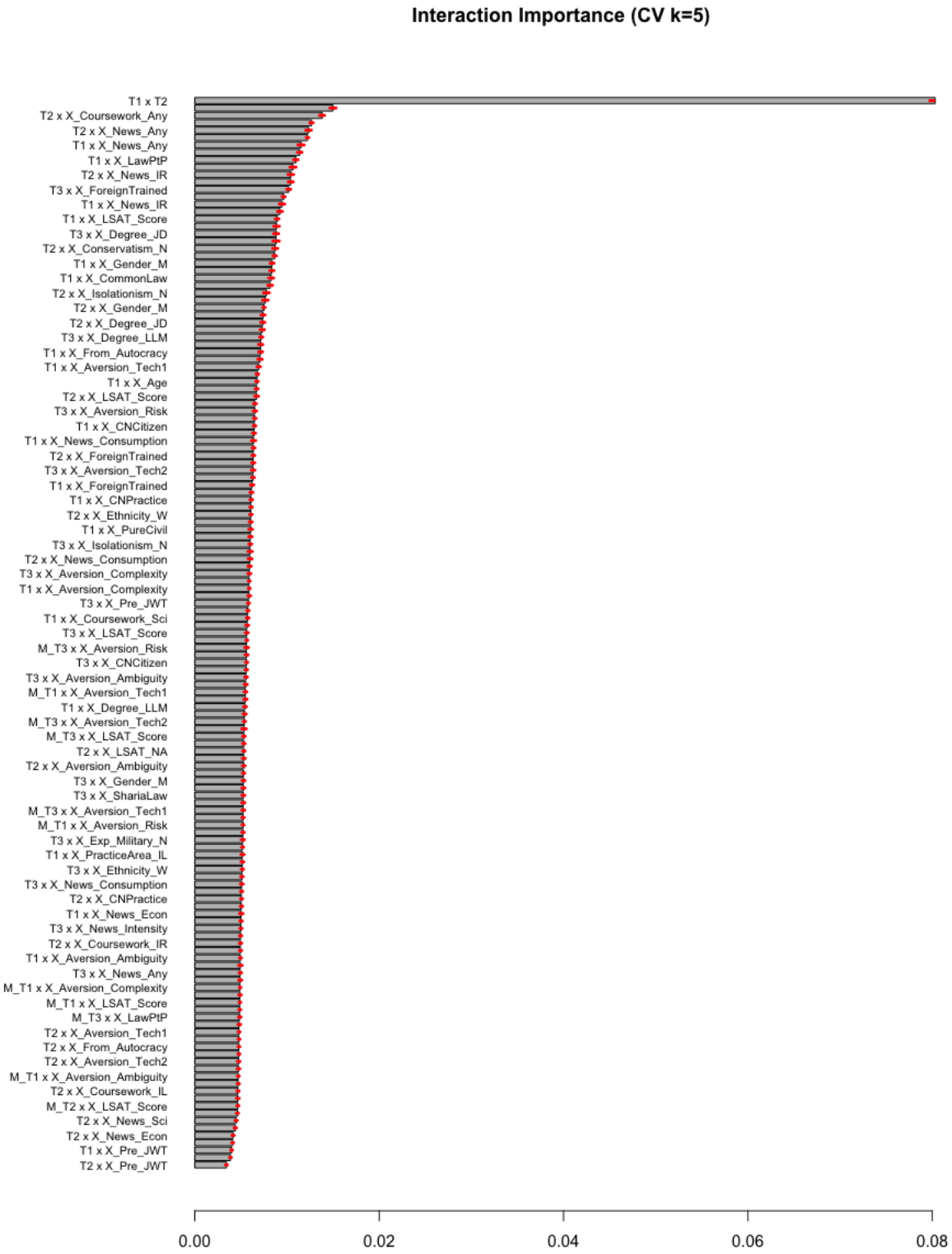
data." I rely on the `bartMachine` package in R to explore all 2-way treatment-by-covariate interactions for a total of 121 variable combinations.

Consistent with best practices, the outcome variable is first scaled between $[-0.5, 0.5]$. This limits BART's sensitivity to different hyperparameter options. Then the sample is randomly divided into training/test sets (90/10). Next, BART is performed with 20-fold cross-validation on a set of six tree configurations ranging between 10 and 200. Dummy variables are automatically included for any covariates with missing data, such as LSAT Score and the five late-addition aversion measures. I use the `investigate_var_importance()` function to obtain variable inclusion proportions for each covariate. This information describes the relative influence of each covariate. I then use the `inflection` package to extract the subset of covariates with the most influence on the outcome.

Using these highly-predictive covariates, I use `bart_machine_get_posterior()` to generate draws from the posterior tree distribution. From this, I obtain posterior credible intervals for all potentially influential treatment-by-covariate interactions. Unlike linear models, these intervals reflect uncertainty in estimating the effect across a diverse range of decision trees. Results should not be interpreted linearly. A negative mean does not imply a strictly negative slope. Instead, it is an average effect derived from partial dependence calculations. Partial dependence in a BART context captures the average change in the predicted response when a specific variable changes, holding other covariates constant. A significant effect implies a general relationship between the variable and the outcome, even if that relationship may not apply uniformly across the entire parameter space.

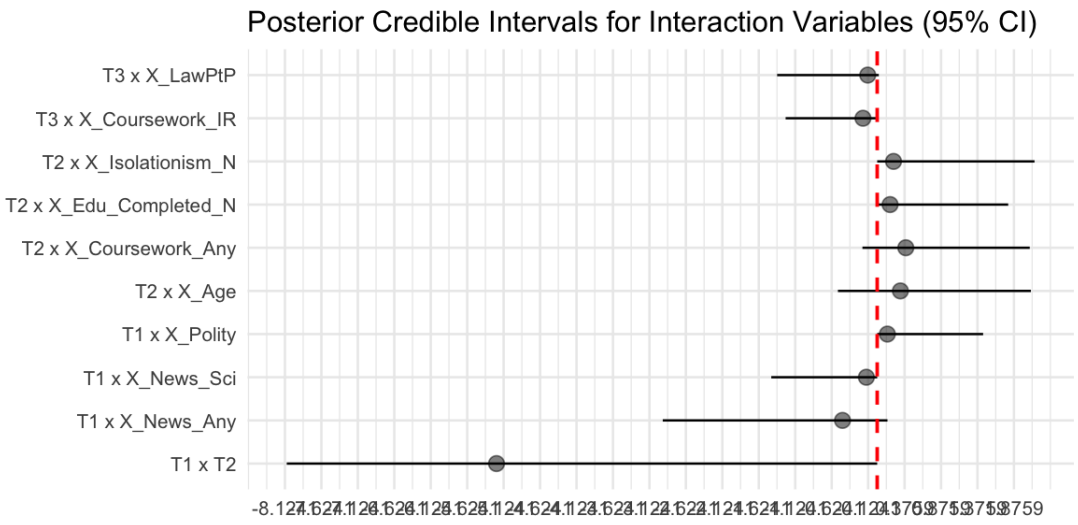
Figure A.8 contains a plot of the inclusion proportions with 95% confidence intervals for the subset of maximally influential treatment-by-covariate candidates. Interaction influence drops off sharply after $T_1 \times T_2$. Since the effect of the main interaction $T_1 \times T_2$ is orders of magnitude larger than the next most influential treatment-by-covariate interactions, the plot is censored after 0.02 to magnify differences between the latter. Partial dependence

Figure A.8: Treatment × Covariate Interaction Search (BART)



Note: Inclusion Proportion plot. Heterogeneous effects below $T_1 \times T_2$ are minute.

Figure A.9: Posterior Means from BART Model (95% Credible Intervals)



Note: Only $T_1 \times T_2$ shows a credible effect.

calculations reflect a high degree of uncertainty and relatively low effect size for all covariate interactions, suggesting that even the likeliest heterogeneous effects are substantively meaningless—too small to be of theoretical interest.

While there may be minute differences in how individuals with certain characteristics respond to each of the treatments, subgroup effects are substantively inconsequential. To see this, we sample from the posterior distribution of fitted effects via Markov chain Monte Carlo (MCMC). For each interaction term, the posterior mean represents the average estimated effect across all posterior draws, while the credible interval spans the central 95% of those sampled values, reflecting uncertainty in the posterior distribution. An effect is considered credibly different from zero when this interval excludes zero. Figure A.9 plots results for the five most-predictive interaction terms. Observe that $T1 \times T2$ is only interaction with strong posterior evidence of a nonzero effect. No treatment-by-covariate interaction has a credible effect.

Unlike in traditional linear models, the credible intervals shown here are derived from

posterior draws of the trees within the BART ensemble, reflecting the uncertainty across the many possible configurations of decision trees that explain the data. A negative posterior mean for a variable or interaction does not imply a simple linear decrease in the outcome; rather, it indicates that, on average across posterior samples, higher values of that variable tend to correspond to lower predicted responses. These means can be viewed as average partial dependence effects, summarizing how changes in a given variable influence predictions while holding others constant. Because BART captures complex, nonlinear relationships, the direction and magnitude of these effects can vary across the covariate space—an interaction may have a strong negative influence in some regions but little to no effect elsewhere. Thus, the intervals represent credible evidence of an effect's overall tendency, not a uniform slope, with only the $T1 \times T2$ interaction showing a consistently nonzero average impact.

A number of diagnostics are performed. These include comparing out-of-sample root mean square error (RMSE), checking RMSE and consistency for all tree configurations, assessing Pseudo R^2 sensitivity, testing whether missingness affects the response, and checking model convergence. All diagnostic code is available in the supplementary material for replication purposes. We can conclude that nearly half of the variation is directly explained by the treatment ($R^2 \approx 0.4$), and the other half by idiosyncrasies unrelated to the covariates. Researchers wishing to recover precise minute effect sizes for certain treatment-covariate combinations could use BART or directly import the variables to a linear model for further exploration.

A.4 Theoretical Mechanisms

A.4.1 Structural Topic Models

To study causal mechanisms, I fit three STM models. Best practices are followed. For consistency, I use the `stm` [Roberts et al. 2014] package in R for document preprocessing. I rely on the `preText` package [Denny and Spirling 2017] to perform sensitivity analysis on the

Table A.4: STM Model Summary

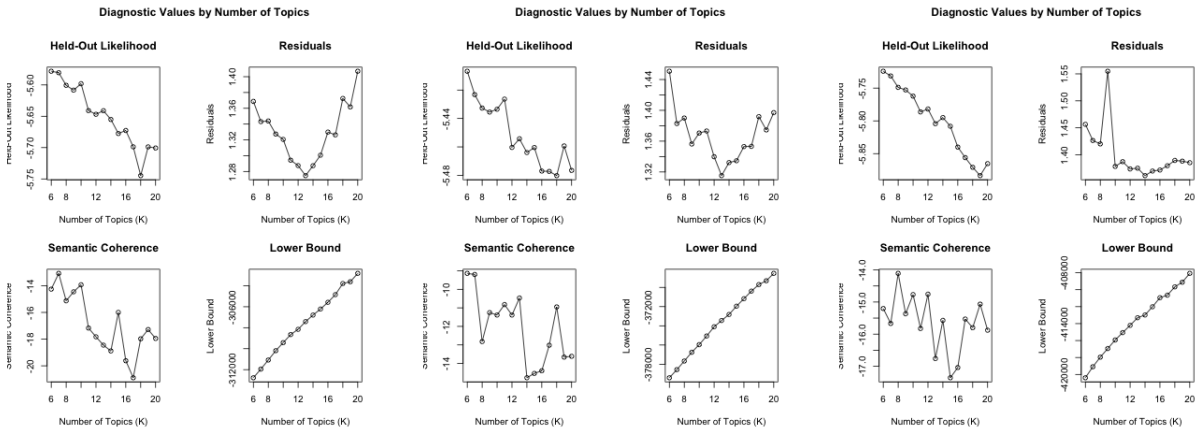
Model	Corpus Subset	Prevalence	K
1	Fixed $T_2 = 1$	$T_1 \times T_3$	9
2	Fixed $T_1 = 1$	$T_2 \times T_3$	9
3	Fixed $Y > 0$	$T_1 \times T_2 \times T_3$	12

full document set, choosing the hyperparameter options recommended in Figure A.11. An exception is made for n-grams for three reasons. First, `stm` offers no native n-gram support, and alternative packages, such as `quanteda`, produced unstable results. Second, n-grams are much more computationally expensive. Third, the tokens produced were difficult to interpret in this case. Absent a theoretical reason to suspect $>$ 1-grams are important to the results, I confine the analysis to unigrams. I also issue minor spelling corrections, perform word-stemming, remove numbers if unconnected to a unit of measurement (eg. 40 degrees, 700,000 kilometers), and convert the text to lowercase Latin-ASCII. All preprocessing steps are available in the replication code.

The optimal number of topics k is obtained for each model with the `stm::searchK()`. I perform a search over $6 : 20$ possible topics and choose the k^* for each model that maximizes semantic coherence and exclusivity based on a visual inspection of the results. Choosing optimal k was relatively uncontroversial in all cases. After building the models, I inspect topic contents using the `stm::labelTopics()` function. I rely on prevalent tokens (ie. β and FREX) along with domain expertise to produce labels that adequately summarize each topic. FREX was especially helpful, since the top-β tokens are non-exclusive. This is an artifact of preText's recommendation not to trim frequent/infrequent words. Topic ratios along with representative tokens are depicted for each model in Figure **??** (β values only). Models exhibit extremely weak-to-nonexistent between-topic correlation (based on t-SNE with the `stm::topicCorr` function) suggesting good stability.

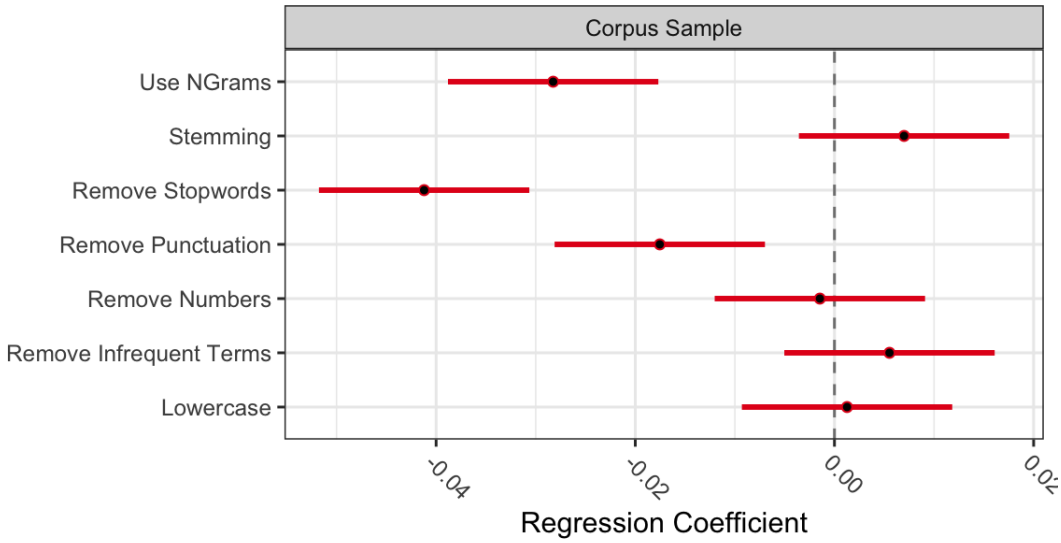
Finally, I estimate the effect of the prevalence covariate shown in Column 3 of Figure A.4 for each model and plot the results using STM's built-in `plot.estimateEffect()` function.

Figure A.10: Search for Optimal Number of Topics (k)



Note: Estimated using `stm` package.

Figure A.11: Hyperparameter Sensitivity Test



Note: Estimated using `preText` package. Negative estimates are recommended.

Plots are included in the main paper in Figure 7. Note the substantial effect sizes. The largest effects in Models 1 and 2 approach -0.4 to $+0.2$. Significant effects in Model 3 are approximately -0.2 to $+0.15$, slightly lower according to the triple interaction term.

A.4.2 Themes and Excerpts

Figure 8 contains paraphrased argumentative excerpts. Full text is available in the replication data. Some additional quotes in support of H_4 and H_5 are included below:

- "It is undeniably the case that XNS presents more of an edge case than any of the precedent satellites" but "this is probably not the type of edge case for which lobbying could make a difference."
- "Other countries will still likely view XNS as restricted, so [our country] should not further invest in XNS to avoid the appearance of noncompliance."
- "Since our country has already reached the 1,000 limit, taking a cautious approach is beneficial. Any noncompliance with the Convention would have a detrimental effect on our countries' [sic] long-term international trade and national security."
- "By developing a satellite that undeniably meets the universally agreed upon definition ... all for the purpose of personal gain, we risk being exposed to the global community as selfish actors [...] Not only will this impair our national credibility, it also has the potential to upend the UN's carefully maintained neutrality of outer space by emboldening other countries to similarly test the outer limits of the Orbital Convention. That is not a reality that we should desire."
- "Our country signed and ratified Orbital Convention because we recognized the importance of avoiding a race to the bottom in humankind's exploration of space. We cannot abandon that long term wisdom for short term gain."
- "Any uncertainty as to whether XNS meets Article 2's criteria works in the favor of

the regulatory authority and to the disadvantage of countries subject to the Orbital Convention."

For added context, Table A.5 contains excerpts (20-word context windows) for specific mentions of "reputation" or "credibility" in participants' written submissions. Observe that such concerns arise regardless of the specificity of the text (T_2) or client orientation (T_3), though lawyers who report to political principals ($T_3 = 1$) appear to find justifications based on legal credibility more persuasive. Excerpts in the paper are drawn from these context windows. Replication code is available in the supplementary material.

Table A.5: Context for Concerns About Noncompliance Exposure

Keyword + Context Window	**Anonymized ID**	**Z**
criteria . If we were to expansively interpret XNS as a permissive satellite , we might jeopardize not only the credibility of our country in the international society , but also the credibility of this Convention that we , along with	JHTLM2513G	T011
in light of a new application is not only unfair to the applicant , but also undermines the authority and credibility of the UN . Conclusion Therefore , I conclude that under the applicable Orbital Convention , XNS is a permitted	NSRLF6436N	T110
as permitted would violate our country's legal obligation , undermine our commitments towards other ratifying parties , jeopardize our national credibility and harm our regional economy . As explained above , XNS meets all technical criteria for being restricted under the	SXVBQ7834T	T011
) . Therefore , an attempt to argue that XNS is not governed by the Convention would jeopardize our nation's credibility B . XNS counts as a " restricted type . " If registered , XNS would need to be counted	IWSQH9089N	T011
neatly into the criteria put forth by the Convention , arguing that XNS is " permitted " would strain our credibility and potentially tarnish our global standing in the field . As general counsel , it is my responsibility to advise	XJHAH3392Z	T101
fealty to a plain-meaning interpretation of the provision . It makes little sense to risk considerable resources and our national credibility for the remote possibility that the UN will disregard an unambiguous provision and its precedent rulings . Beyond the plain-meaning	JLZHZ5225V	T011

as selfish actors in what was intended to be a universalist region . Not only will this impair our national credibility , it also has the potential to upend the UN's carefully maintained neutrality of outer space by emboldening other countries	JLZHZ5225V	T011
assessment . Arbitrarily restricting a satellite that meets none of our requirements for a restricted satellite would severely undermine our credibility as an institution . Doing so would make us into an unpredictable group instead of one that follows the law	UDWVG4082P	T110
the criteria are guidelines rather than strict cut-offs . Such an interpretation would lead to subjective determinations , undermining the credibility and effectiveness of the UN's determinations . Although most of the satellites listed in Schedule 1 (per our prior	XVQKI1465U	T010
realm ? Stretching the terms of the Orbital Convention to make XNS a restricted satellite would not only ruin the credibility of this office , but would seemingly penalize a country for its innovation . If countries have such an adverse	MRQGM2056C	T110
could be enough not to meet the Article's criteria , but considering the country's desire to keep its strong national credibility , this could be considered too permissive a reading . I wouldn't recommend using this criteria as reason for "	OWQWS3941N	T011
, that would make the path to " permitted " classification much easier . However , the country's position and credibility are not to be taken lightly , and it may be best not to push forward with a permissive reading	OWQWS3941N	T011
. Arguing that XNS should be considered a permitted satellite without strong evidence of the satellite's safety could undermine our credibility .	RVPCR4972T	T011
authority , as a body lacking true sovereign enforcement power , is likely to be extremely sensitive to maintaining its credibility , which would be severely undercut if it made an exception for our country . While Article IV of the	QRZNR1006J	T011
already exceeded our maximum restricted satellites , thus putting us in violation of the convention obligation and jeopardizing our national credibility . Lastly , under Article IV we have waived our right to withdraw from the treaty . Thus , rather	DJNHG3526U	T011
Taking a more permissive interpretation of the Orbital Convention in order to move forward with XNS may jeopardize our national credibility .	AQTKA5257O	T011

against the plain meaning . Given our belief , it would be unwise to waste money and jeopardize our national credibility pursuing XNS . XNS meets all the requirements of a restricted satellite . According to Article II of the Orbital	KECXZ6002I	T011
to be classified as a permitted satellite , our country would have to lobby the United Nations , risking our credibility for an outcome we believe to be extremely unlikely . To count XNS as permitted , the United Nations registration	KECXZ6002I	T011
registration authority would be swayed by arguments of this nature , we believe our country should not risk jeopardizing our reputation and wasting government funds on XNS .	KECXZ6002I	T011

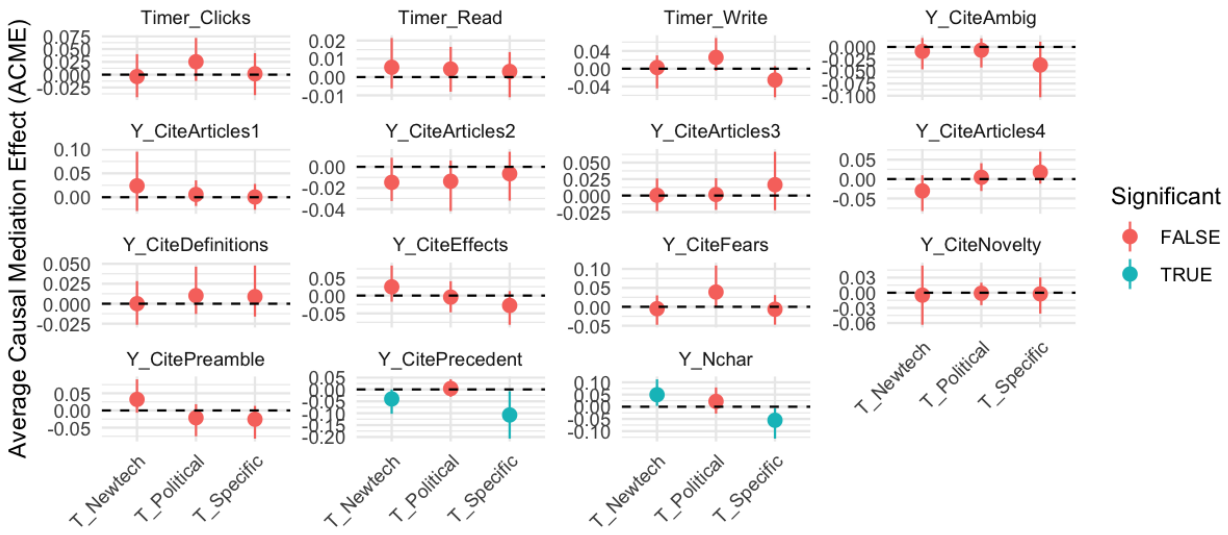
A.4.3 Mediation Analysis

This section considers the possibility of causal mediation on the outcome, perceived applicability of the Orbital Convention [Imai et al. 2010]. Mediation analysis seeks to understand the pathways (M) through which treatment assignment (Z, equivalently D with full compliance) can affect the outcome (Y). It distinguishes between direct effects ($Z \to Y$), indirect effects ($M \to Y$), and total effect ($Z \to M \to Y$). First, relationships are decomposed by running three separate models. Second, following Hicks and Tingley [2011], a full causal mediation model is implemented using the `mediation` package in R [Tingley et al. 2014]. I treat this exercise as exploratory, as it was not preregistered. Code is available in the supplementary material.

Fifteen potential mediators are considered, ranging from what participants cited in their written briefs to their level of indecision (proxied by click rates and time spent reading the prompt). Potential mediators included `CiteAmbig`, `CiteArticles1`, `CiteArticles2`, `CiteArticles3`, `CiteArticles4`, `CitePrecedent`, `CitePreamble`, `CiteNovelty`, `CiteEffects`, `CiteDefinitions`, `CiteFears`, `Nchar`, `Timer_Read`, `Timer_Write`, and `Timer_Clicks`. Average Causal Mediation Effects (ACMEs), Average Direct Effects

(ADEs), and total effects were estimated for each potential mediator in the full factorial design. For every mediator, I fit linear models of the mediator and the outcome on the three experimental factors (`T_Newtech`, `T_Specific`, `T_Political`) and the `T_Newtech` × `T_Specific` interaction. I then computed nonparametric bootstrapped confidence intervals based on 1,000 simulations.

Figure A.12: Significant Mediation (ACME)



Note: Only ACMEs are shown.

Figure A.12 plots significant ACMEs. Across all fifteen mediators and three treatments, nearly all ACME intervals overlapped zero, indicating no statistically reliable indirect effects. Only four mediator-treatment pairs showed confidence intervals that excluded zero: `Y_Nchar` for the `T_Specific` treatment ($\widehat{\text{ACME}} = -0.055$, 95% CI $[-0.126, -0.004]$), `Y_Nchar` for the `T_Novel` treatment ($\widehat{\text{ACME}} = -0.001$, 95% CI $[-0.049, 0.001]$), `Y_CitePrecedent` for the `T_Novel` treatment ($\widehat{\text{ACME}} = 0.0399$, 95% CI $[-0.103, 0.0002]$), and `Y_CitePrecedent` for the `T_Specific` treatment ($\widehat{\text{ACME}} = -0.108$, 95% CI $[-0.223, -0.007]$). The first, second, and third of these is small and negative, suggesting that when participants were exposed to

the `Specific` treatment, they cited slightly fewer precedents and produced shorter responses, and when exposed to the novel technology treatment, they cited less precedent. Conversely, when deliberating over a novel technology, they tended to write slightly more. These results are illustrative but unsuprising. In all cases, the direct effects captured most of the total treatment influence. Overall, I find little evidence that the measured textual or timing variables mediate the effects of the factorial treatments on participants' judgments; the indirect pathways, if present, appear weak and statistically indistinguishable from zero.

A.5 Codebook

A.5.1 Derivative Covariates

As discussed, 15 covariates are measured in a pretreatment questionnaire. Five additional questions were added for the last 20% of the sample. The latter are intended to measure different types of aversion (risk, complexity, ambiguity, and two questions about technology). Aversion covariates are excluded from the analysis but could be informally illustrative. Additional covariates are derived from observed responses to these questions as follows:

1. `X_USCitizen`: Derived from `X_Location_From`, a question about country of citizenship. Coded 1 if from the United States or outlying territories; 0 otherwise.
2. `X_CNCitizen`: Derived from `X_Location_From`, a question about country of citizenship. Coded 1 if from mainland China, Hong Kong, or Macau; 0 otherwise.
3. `X_USPractice`: Derived from `X_Location_Work`, a question about intended or actual jurisdiction of professional practice. Coded 1 if US jurisdiction; 0 otherwise.
4. `X_CNPractice`: Derived from `X_Location_Work`, a question about intended or actual jurisdiction of professional practice. Coded 1 if mainland China jurisdiction; 0 otherwise.

5. `X_CommonLaw`: Derived from `X_Location_From`. Country of origin is matched with crossnational legal tradition data from Guerriero [2016]. Coded 1 if origin country possesses common law attributes.

6. `X_ShariaLaw`: Derived from `X_Location_From`. Country of origin is matched with crossnational legal tradition data from Guerriero [2016]. Coded 1 if origin country possesses sharia (religious) law attributes.

7. `X_PureCivil`: Derived from `X_Location_From`. Country of origin is matched with crossnational legal tradition data from Guerriero [2016]. Coded 1 if origin country possesses no common or sharia law attributes, only pure civil law.

8. `X_Coursework_IL`: Derived from `X_Coursework`, which asks participants about the types of higher education courses they have undertaken. Coded 1 if a participant reported taking coursework in international, national security, or technology law; 0 otherwise.

9. `X_Coursework_IR`: Derived from `X_Coursework`, which asks participants about the types of higher education courses they have undertaken. Coded 1 if a participant reported taking coursework in international relations; 0 otherwise.

10. `X_Coursework_Sci`: Derived from `X_Coursework`, which asks participants about the types of higher education courses they have undertaken. Coded 1 if a participant reported taking coursework in physics, astronomy, or engineering; 0 otherwise.

11. `X_Coursework_Any`: Coded 1 if a participant reported any of the above courses; 0 otherwise.

12. `X_Degree_JD`: Derived from `X_Degree`. Coded 1 if law degree is a three-year advanced juris doctorate (JD); 0 otherwise.

13. `X_Degree_LLM`: Derived from `X_Degree`. Coded 1 if law degree is a one-year advanced Master of Laws (LLM) degree; 0 otherwise. Students must possess a professional law

degree, such as a Bachelors of Law (LLB), in order to pursue an LLM.

14. `X_Ethnicity_W`: Derived from `X_Ethnicity`, a multiple choice factor variable. Coded 1 if a participant self-identified at least partly as White; 0 otherwise.

15. `X_ForeignTrained`: Derived from `X_Degree`. Coded 1 if obtained an LLB or LLM degree.

16. `X_Gender_M`: Derived from `X_Gender`. Coded 1 if a participant self-identified as male; 0 otherwise.

17. `X_LSAT_NA`: Derived from `X_LSAT_Score`. Coded 1 if a participant did not take the Law School Admissions Test (LSAT); 0 otherwise.

18. `X_News_Econ`: Derived from `X_NewsType`, a question about the type of news participants typically consume. Coded 1 if related to economics, finance, or business; 0 otherwise.

19. `X_News_IR`: Derived from `X_NewsType`, a question about the type of news participants typically consume. Coded 1 if related to international affairs, foreign policy, or national security; 0 otherwise.

20. `X_News_Sci`: Derived from `X_NewsType`, a question about the type of news participants typically consume. Coded 1 if related to science, technology, or outer space; 0 otherwise.

21. `X_News_Any`: Coded 1 if any of the above was reported; 0 otherwise.

22. `X_News_Intensity`: Derived by multiplying `X_NewsFrequency` (5-point scale) with `X_News_Any`.

23. `X_Polity`: Derived by matching `X_Location_From` (country of origin) with that country's Polity V score.[25]

[25] https://www.systemicpeace.org/polityproject.html

24. `X_From_Autocracy`: Derived by matching `X_Location_From` (country of origin) with that country's Polity V score. Coded 1 if country has a negative Polity score; 0 otherwise.

25. `X_PracticeArea_IL`: Derived from `X_PracticeArea`, a question about actual or intended practice specialization. Coded 1 if includes international, national security, or technology law; 0 otherwise.

26. `X_Practitioner`: Derived from `X_EduCompleted`, a question about the number of years of training a participant has completed, up to the maximum (four and beyond). Coded 1 if a participant reported having reached 3L status (the point at which outside clinical training commences) or has graduated; 0 otherwise.

27. `X_LawPtP`: Proxies for legal bureaucratization at the country-level. Derived from biographical data on cabinet-level officials compiled by Nyrup et al. [2025]. Expressed as the proportion of officials with terminal educational degrees in law. Proportions for each country are matched to `X_Location_From`.

28. `X_Pre_JWT`: A dummy variable indicating whether a participant completed the study prior to the James Webb Space Telescope launch in December 2021.

29. `X_WaveOrder`: A continuous variable indicating order of participation for time trends purposes.

30. `X_Semester`: A factor variable indicating wave timing by recruitment announcement date.

31. `X_BatchID`: A factor variable indicating the specific NSLS $\times$ semester delivery network.

32. `X_LawSchoolID`: An indicator for law school affiliation (anonymized).

A.5.2 Secondary Outcomes

Secondary outcomes are derived from survey metadata, including written responses. The distributions for six of these outcomes are plotted in Figure A.5:

1. `Y_Likert`: After receiving treatment, participants are asked for a legal opinion on whether the Orbital Convention applies (Yes/No/Not Sure) and to report their confidence (3-point scale). These are multiplied together to obtain a 7-point likert scale, which is used as the main dependent variable.

2. `Y_CitePramble`: Derived from `Y_Articles`, a question about the Orbital Convention provisions that a participant thinks are most relevant to the ruling, and `Y_Essay`, the long-form brief. Coded 1 if participant cited the Preamble; 0 otherwise.

3. `Y_CiteArticles1`: Derived from `Y_Articles` and `Y_Essay`. Coded 1 if participant cited Article I (basic obligation, distinction between permitted and restricted satellites); 0 otherwise.

4. `Y_CiteDefinitions`: Derived from `Y_Articles` and `Y_Essay`. Coded 1 if participant cited Article II (four operative criteria); 0 otherwise.

5. `Y_CiteArticles3`: Derived from `Y_Articles` and `Y_Essay`. Coded 1 if participant cited Article III (bindingness); 0 otherwise.

6. `Y_CiteArticles4`: Derived from `Y_Articles` and `Y_Essay`. Coded 1 if participant cited Article IV (procedure); 0 otherwise.

7. `Y_CitePrecedent`: Derived from `Y_Articles` and `Y_Essay`. Coded 1 if participant cited Schedule I (past satellite classifications); 0 otherwise.

8. `Y_CiteFears`: Derived from `Y_Articles` and `Y_Essay`. Coded 1 if participant cited keywords related to allegation-avoidance or credibility; 0 otherwise.

9. `Y_CiteNovelty`: Derived from `Y_Articles` and `Y_Essay`. Coded 1 if participant cited

keywords related to technological novelty; 0 otherwise.

10. `Y_CiteAmbig`: Derived from `Y_Articles` and `Y_Essay`. Coded 1 if participant cited keywords related to specificity or ambiguity; 0 otherwise.

11. `Y_CiteEffects`: Derived from `Y_Articles` and `Y_Essay`. Coded 1 if participant cited keywords related to the societal or material impact of the technology; 0 otherwise.

12. `Y_Nchar`: Character length of `Y_Essay`.

13. `Y_Nwords`: Word count of `Y_Essay`.

14. `Timer_Read`: Derived from a Qualtrics indicator for how many seconds a participant spent reading an introductory vignette immediately prior to treatment assignment.

15. `Timer_Write`: Derived from a Qualtrics indicator for how many seconds a participant spent on the long-form writing portion.

16. `Timer_Clicks`: Derived from a Qualtrics indicator for how many times a participant clicked a mouse button during the long-form writing portion. A proxy for the number of edits/revisions.

A.6 Appendix Bibliography

Imai, Kosuke, Luke Keele, and Dustin Tingley (2010). "A general approach to causal mediation analysis". In: *Psychological Methods* 15.4. Place: US Publisher: American Psychological Association, pp. 309–334.

Hicks, Raymond and Dustin Tingley (Dec. 1, 2011). "Causal Mediation Analysis". In: *The Stata Journal* 11.4. Publisher: SAGE Publications, pp. 605–619.

Green, Donald P. and Holger L. Kern (Jan. 1, 2012). "Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees". In: *Public Opinion Quarterly* 76.3, pp. 491–511.

Gelman, Andrew and John Carlin (Nov. 1, 2014). "Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors". In: *Perspectives on Psychological Science* 9.6. Publisher: SAGE Publications Inc, pp. 641–651.

Roberts, Margaret E. et al. (2014). "Structural Topic Models for Open-Ended Survey Responses". In: *American Journal of Political Science* 58.4. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12103, pp. 1064–1082.

Tingley, Dustin et al. (Sept. 2, 2014). "mediation: R Package for Causal Mediation Analysis". In: *Journal of Statistical Software* 59, pp. 1–38.

Guerriero, Carmine (May 31, 2016). "A novel dataset on legal traditions, their determinants, and their economic role in 155 transplants". In: *Data in Brief* 8, pp. 394–398.

Denny, Matthew and Arthur Spirling (Sept. 27, 2017). *Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What/Do about It*. SSRN Scholarly Paper ID 2849145. Rochester, NY: Social Science Research Network.

Brutger, Ryan and Joshua D. Kertzer (2018). "A Dispositional Theory of Reputation Costs". In: *International Organization* 72.3. Publisher: Cambridge University Press, pp. 693–724.

Montgomery, Jacob M., Brendan Nyhan, and Michelle Torres (2018). "How Conditioning on Posttreatment Variables Can Ruin Your Experiment and What/Do about It". In: *American Journal of Political Science* 62.3, pp. 760–775.

Baranger, David A. A. et al. (July 1, 2023). "Tutorial: Power Analyses for Interaction Effects in Cross-Sectional Regressions". In: *Advances in Methods and Practices in Psychological Science* 6.3. Publisher: SAGE Publications Inc, p. 25152459231187531.

Muralidharan, Karthik, Mauricio Romero, and Kaspar Wüthrich (Mar. 15, 2023). "Factorial Designs, Model Selection, and (Incorrect) Inference in Randomized Experiments". In: *The Review of Economics and Statistics*, pp. 1–44.

Sommet, Nicolas et al. (July 1, 2023). "How Many Participants Do I Need to Test an Interaction? Conducting an Appropriate Power Analysis and Achieving Sufficient Power to Detect an Interaction". In: *Advances in Methods and Practices in Psychological Science* 6.3. Publisher: SAGE Publications Inc, p. 25152459231178728.

Nyrup, Jacob et al. (2025). "Paths to Power: A New Dataset on the Social Profile of Governments". In: *British Journal of Political Science* 55, pp. 1–20.