Online Appendix

Model Initialization and Configuration

To evaluate how structural features of the interaction environment shape opinion dynamics, we vary both the initial distribution of opinions and key parameters of the simulated population. We design the model context to reflect empirically observed opinion distributions. Using survey data from the 2020 American National Election Study (ANES) (American National Election Studies, 2021), we calculate the mean and standard deviation of self-reported ideological placements for respondents in each U.S. state. Among the state-level parameters, we retrieve the minimum and maximum levels of both mean ideological placement and standard deviation and use them to create four bimodal opinion distributions, with combinations of moderate versus extreme opinions and low versus high opinion variation. Agents' initial opinions are drawn randomly from these distributions.

Table 4 presents the means and standard deviations of each type of opinion distribution, scaled to be bounded on the $[-1, 1]$ interval. Agents' initial opinions are drawn from a normal distribution parameterized by these statistics. These distributions allow us to assess how contextual differences in the starting configuration of attitudes influence subsequent opinion dynamics.

Table 4: Initial Opinion Distributions

	Mean	**Std. Dev.**
Moderate opinion, low variation	0.05	0.225
Moderate opinion, high variation	0.05	0.3976
Extreme opinion, low variation	-0.4242	0.225
Extreme opinion, high variation	-0.4242	0.3976

Population size varies across $\{500, 600, 700, 800, 900, 1000\}$, resulting in different density conditions. Agents are connected through a small-world network, where the average neighborhood size varies across $\{4, 5, 6, 7, 8\}$. While we control the mean degree, the small-world topology permits substantial variation in individual degrees, with some agents having many more ties than others. Larger average neighborhood sizes increase both the likelihood of encountering friends and the probability of higher learning rates (since friend interactions double the learning rate). Each combination of neighborhood size and population size are evaluated in the context of each of the four initial opinion distributions.

A.1 Sensitivity analyses

A.1.1 Bayesian updating under the normally distributed opinion environment

Table 5 reports the results for the sensitivity analysis in which the initial opinion distribution follows a normal rather than a uniform structure. The model again evaluates two dimensions of opinion dynamics: the Camp-Switching Rate (CSR) and the Mean Absolute Change (MAC).

The core patterns observed in the baseline Bayesian models remain mostly intact. Interaction exerts a strong positive effect on both ideological switching and the magnitude of opinion change. In the baseline context (moderate opinion with low variation and no elite intervention), higher interaction increases the probability of camp switching ($\beta = 0.039$, $p < 0.001$) and also increases the magnitude of opinion displacement ($\beta = 0.033$, $p < 0.001$). This indicates that even under a normally distributed opinion landscape, social encounters function as informational signals that encourage agents to revise their beliefs.

The structural configuration of the opinion environment again shapes these dynamics. Relative to the baseline context, extreme environments suppress ideological switching. Both the extreme–high variation ($\beta = -0.333$, $p < 0.001$) and extreme–low variation ($\beta = -0.347$, $p < 0.001$) contexts reduce the likelihood that agents cross the neutral ideological threshold. At the same time, these environments generate larger absolute opinion movements. For example, the extreme–high variation context increases MAC by 0.212 units ($p < 0.001$). This means that agents in polarized environments move within ideological camps.

Elite cues matter. Positive elite signals increase both camp switching ($\beta = 0.022$, $p < 0.001$) and the magnitude of opinion displacement ($\beta = 0.040$, $p < 0.001$), suggesting that agents respond strongly when elite communication points toward a particular ideological direction. In contrast, negative elite cues slightly reduce the probability of ideological switching ($\beta = -0.012$,

$p < 0.001$) and have no statistically significant effect on the magnitude of opinion change. These findings indicate that elite messaging can bias the direction of opinion movement but does not fundamentally alter the underlying updating process.

The interaction terms reveal that the effect of interaction depends on the surrounding opinion environment. In extreme contexts, interaction substantially weakens the probability of camp switching. For example, the interaction between interaction frequency and the extreme–high variation environment reduces CSR by 0.081 ($p < 0.001$). At the same time, these environments amplify the magnitude of within-camp opinion movement, as indicated by the large positive interaction effect for MAC in the extreme–high variation context ($\beta = 0.147$, $p < 0.001$). This pattern again suggests that interaction in polarized environments tends to intensify existing ideological positions rather than produce ideological realignment.

Neighborhood size also exhibits a stabilizing effect across both outcomes. Larger interaction neighborhoods reduce both ideological switching ($\beta = -0.006$, $p < 0.001$) and the magnitude of opinion change ($\beta = -0.012$, $p < 0.001$), indicating that agents embedded in larger local networks are less responsive to individual encounters. Overall, these results confirm that the core dynamics observed in the baseline Bayesian models are robust to alternative assumptions about the initial distribution of opinions.

A.1.2 The Backlash mechanism under the normally distributed opinion environment

Table 6 reports the results of two regression models examining how interaction frequency, opinion context, and elite signaling shape opinion dynamics under the backlash mechanism. Overall, the results indicate that while interaction can increase the frequency of opinion updates, its effect depends mainly on the surrounding opinion environment and by elite cues. The backlash mechanism, therefore, produces opinion change patterns that are highly contingent on contextual structure.

In the baseline context—moderate opinion with low variation—interaction increases both the frequency and magnitude of opinion change. The coefficient for Interaction (Scaled) is positive and statistically significant in both models (CSR: $\beta = 0.015$, $p < 0.001$; MAC: $\beta = 0.140$, $p <$

Table 5: Regression Results: Bayesian Updating (Normal Distribution + Elite)

	Dependent Variable	
	Camp-Switching Rate (CSR)	Mean Absolute Change (MAC)
Interaction (Scaled)	0.039^{***}	0.033^{***}
	(0.003)	(0.003)
Context: Extreme (High Var)	-0.333^{***}	0.212^{***}
	(0.006)	(0.007)
Context: Extreme (Low Var)	-0.347^{***}	0.070^{***}
	(0.009)	(0.009)
Context: Moderate (High Var)	-0.113^{***}	-0.023^{***}
	(0.003)	(0.003)
Elite: Negative (-1)	-0.012^{***}	-0.002
	(0.001)	(0.001)
Elite: Positive (+1)	0.022^{***}	0.040^{***}
	(0.001)	(0.001)
Neighborhood Size	-0.006^{***}	-0.012^{***}
	(0.0002)	(0.0002)
Interaction × Ext (High Var)	-0.081^{***}	0.147^{***}
	(0.007)	(0.007)
Interaction × Ext (Low Var)	-0.063^{***}	0.027^{**}
	(0.009)	(0.009)
Interaction × Mod (High Var)	-0.023^{***}	-0.010^{**}
	(0.003)	(0.003)
Constant	0.392^{***}	0.177^{***}
	(0.003)	(0.003)
Observations	5,400	5,400
R^2	0.937	0.565
Adjusted R^2	0.937	0.564
Residual SE	0.038 (df = 5389)	0.039 (df = 5389)
F Statistic	$8{,}025.036^{***}$ (df = 10; 5389)	700.030^{***} (df = 10; 5389)

Note: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

0.001). This suggests that greater social exposure promotes opinion updating when the surrounding environment contains relatively moderate viewpoints.

However, the interaction terms reveal that this effect varies across opinion contexts. In extreme environments, increased interaction dampens opinion change. The interaction between interaction frequency and extreme contexts is negative and statistically significant in both cases (e.g., Interaction $\times$ Extreme High Variation: CSR $\beta = -0.015$, MAC $\beta = -0.121$, $p < 0.001$). Substantively, this indicates that encounters with highly polarized viewpoints trigger backlash responses that suppress further updating. In contrast, in the moderate high-variation context, interaction produces a more nuanced pattern. While interaction increases the rate of camp switching (CSR interaction coefficient $= 0.028$, $p < 0.001$), it simultaneously reduces the magnitude of opinion change (MAC interaction coefficient $= -0.046$, $p < 0.001$). In other words, agents switch camps more frequently but do so through relatively small opinion adjustments. This pattern suggests that moderate diversity encourages incremental movement rather than large ideological shifts.

Independent of interaction frequency, the overall opinion context influences opinion dynamics. Relative to the baseline context of moderate opinions with low variation, all other contexts reduce camp switching. The most pronounced effect appears in extreme low-variation environments ($\beta = -0.363$, $p < 0.001$), where ideological homogeneity around extreme positions limits movement between camps. A similar pattern emerges for the magnitude of opinion change. Both extreme contexts reduce MAC relative to the baseline, although the magnitude of this effect varies across distributions. These findings suggest that when populations begin with extreme opinion structures, the backlash mechanism tends to stabilize those positions rather than encourage convergence.

Elite cues matter as well. Negative elite signals—which reinforce partisan hostility—reduce both the frequency and magnitude of opinion change (CSR: $\beta = -0.017$, MAC: $\beta = -0.027$, $p < 0.001$). In contrast, positive elite signals slightly increase both outcomes (CSR: $\beta = 0.013$, MAC: $\beta = 0.012$, $p < 0.001$). Although the substantive magnitudes are smaller than the contextual effects, the direction of these coefficients suggests that elite messaging can either dampen or facilitate opinion adjustment depending on the tone of the signal.

Finally, neighborhood size exerts a consistent negative effect across both models. Larger local networks reduce both the probability of camp switching ($\beta = -0.005$, $p < 0.001$) and the magnitude of opinion change ($\beta = -0.027$, $p < 0.001$). Within the backlash framework, larger interaction neighborhoods likely increase exposure to disagreement, which, in turn, triggers defensive responses that stabilize existing attitudes.

The models explain a substantial proportion of the variance in opinion dynamics. The CSR model achieves an R^2 of 0.955, indicating that the predictors capture nearly all variation in camp switching across simulation runs. The MAC model explains a smaller but still substantial share of variation ($R^2 = 0.656$), suggesting that the magnitude of opinion movement is somewhat less predictable than the frequency of switching.

Overall, the results demonstrate that under the backlash mechanism, interaction does not uniformly promote opinion convergence. Instead, the effect of interaction depends critically on the surrounding opinion distribution and the presence of elite cues. In moderate environments, interaction encourages opinion updating, whereas in polarized contexts it can suppress change by triggering defensive responses. These findings highlight the importance of contextual structure in shaping the aggregate consequences of social interaction.

A.1.3 Willingness to talk

Tables 7 and 8 report results from two sensitivity analyses that vary agents' baseline willingness to engage in cross-cutting conversation. We test whether the model's findings are robust to this parameterization.

In Table 7 (willingness threshold at 5%), the interaction ratio exerts a negative effect on both CSR ($\beta = -0.261$, $p < 0.001$) and MAC ($\beta = -0.513$, $p < 0.001$), a reversal from the main models. This pattern reflects a ceiling effect: when agents are initialized with very low willingness to engage across lines of difference, the baseline rate of opinion movement is high relative to the marginal contribution of additional encounters. This, in turn, crowds out the informational diversity needed to sustain updating. Consistent with the main findings, neighborhood size de-

Table 6: Regression Results: Backlash Mechanism (Normal Distribution + Elite)

	Dependent Variable	
	Camp Switching Rate (CSR)	Mean Absolute Change (MAC)
Interaction (Scaled)	0.015^{***}	0.140^{***}
	(0.001)	(0.002)
Context: Extreme (High Var)	-0.334^{***}	-0.087^{***}
	(0.001)	(0.003)
Context: Extreme (Low Var)	-0.363^{***}	-0.007^{*}
	(0.001)	(0.003)
Context: Moderate (High Var)	-0.081^{***}	-0.101^{***}
	(0.001)	(0.003)
Elite: Negative (-1)	-0.017^{***}	-0.027^{***}
	(0.001)	(0.003)
Elite: Positive (+1)	0.013^{***}	0.012^{***}
	(0.001)	(0.003)
Neighborhood Size (n-size)	-0.005^{***}	-0.027^{***}
	(0.0002)	(0.0005)
Interaction $\times$ Ext (High Var)	-0.015^{***}	-0.121^{***}
	(0.001)	(0.004)
Interaction $\times$ Ext (Low Var)	-0.015^{***}	-0.106^{***}
	(0.001)	(0.004)
Interaction $\times$ Mod (High Var)	0.028^{***}	-0.046^{***}
	(0.001)	(0.003)
Constant	0.426^{***}	0.509^{***}
	(0.002)	(0.004)
Observations	5,400	5,400
R^2	0.955	0.656
Adjusted R^2	0.955	0.655
Residual SE	0.036 (df = 5389)	0.084 (df = 5389)
F Statistic	$11{,}341.290^{***}$ (df = 10; 5389)	$1{,}027.752^{***}$ (df = 10; 5389)

Note: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

presses both outcomes, and elite messaging retains a positive effect—positive cues increase CSR and MAC, and, notably, negative elite cues also produce small but significant positive effects on both outcomes, suggesting that any directional signal raises the informational temperature of the environment when willingness is otherwise constrained.

The interaction terms in Table 7 restore the contextual moderation pattern observed in the main models. In extreme high-variation environments, the negative effect of interaction is offset ($\beta = 0.198$ for CSR; $\beta = 0.572$ for MAC), and in moderate low-variation contexts the offset is even larger for CSR ($\beta = 0.240$). This implies that low willingness suppresses interaction-driven updating least in contexts with high opinion diversity. In such settings, polarized and clustered environments provide more cross-cutting opportunities, reducing insulation. For the extreme low-variation context, the interaction term for MAC is small and statistically insignificant ($\beta = -0.037$), indicating that the most polarized scenario resists moderation regardless of encounter frequency. Model fit is very high ($R^2 = 0.990$ for CSR; $R^2 = 0.892$ for MAC) because the data come from controlled experiments run through simulations.

Table 8 raises the willingness threshold to 20%. The negative main effect of interaction ratio persists but attenuates for CSR ($\beta = -0.148$) and becomes negligible for MAC ($\beta = -0.031$). This suggests partial convergence with the main results as willingness increases. The most striking feature of Table 8 is the magnitude of the MAC coefficient for extreme high-variation contexts ($\beta = 1.240$, $p < 0.001$) and its corresponding interaction term ($\beta = 1.257$, $p < 0.001$). Under higher willingness, polarized environments produce larger opinion movements when interaction is more frequent. This pattern is consistent with the theoretical expectation that cross-cutting encounters in ideologically distant populations carry substantial updating potential. The CSR pattern for these same contexts is nonetheless negative ($\beta = -0.346$), which reinforces the established asymmetry between movement magnitude and directional crossing: interaction in polarized settings amplifies within-camp opinion displacement without reliably producing camp-switching. Moderate low-variation contexts switch sign relative to Table 7, with both CSR ($\beta = 0.108$) and MAC ($\beta = 0.046$) showing positive effects, confirming that higher willingness restores the depolarizing potential of interaction in less extreme environments. Model fit remains high ($R^2 = 0.979$ for CSR; $R^2 = 0.967$ for MAC) given that we use simulation data.

Overall, interaction alone is not sufficient to generate opinion change. As Tables 7 and 8 show that the effects of interaction depend on whether individuals are willing to engage with disagreement in the first place. When that willingness is limited, increasing the volume of interaction amplifies within-group reinforcement rather than cross-cutting persuasion. Thus, agents reinforce their existing views rather than updating them.

Table 9 examines whether the learning rate multiplier applied to strong-tie interactions—specifically, whether doubling (2.5×) versus moderately boosting (1.5×) the learning rate for friend dyads—alters opinion dynamics. The coefficient for the friend-learning indicator is effectively zero and statistically indistinguishable from null for both CSR ($\beta = -0.0001$) and MAC ($\beta = -0.0002$). This null result indicates that the magnitude of the trust assigned to network friends does not independently shape aggregate opinion dynamics, once the structural properties of the interaction environment are controlled. The influence of strong ties on opinion formation operates through the network topology itself rather than through fine-grained variation in tie strength.

All other results in Table 9 replicate the established pattern in the main results. Interaction ratio is negative for both outcomes in the baseline context ($\beta = -0.275$ for CSR; $\beta = -0.318$ for MAC), neighborhood size is consistently stabilizing, and elite cues increase both CSR and MAC. The interaction of interaction ratio with extreme high-variation contexts again yields a large positive effect on MAC ($\beta = 0.875$, $p < 0.001$), while the corresponding CSR term is also positive and significant ($\beta = 0.182$). By contrast, the extreme low-variation interaction terms are small and do not reach conventional significance thresholds for either outcome, echoing the resistance to updating observed in the most polarized configuration across model variants. Elite positive cues increase both outcomes at comparable magnitudes to prior models (CSR: 0.027; MAC: 0.046), and elite negative cues retain small positive effects. The model achieves high explanatory power ($R^2 = 0.985$ for CSR; $R^2 = 0.885$ for MAC). Overall, these results confirm that the core dynamics reported in the main text are robust to variation in the trust-weighting of social ties.

Table 7: Regression Results: Camp-Switching Rate and Mean Absolute Change (Willingness (5%))

	Dependent variable:	
	Camp-Switching Rate (CSR)	Mean Absolute Change (MAC)
	CSR	MAC
Interaction Ratio (scaled)	-0.261^{***}	-0.513^{***}
	(0.004)	(0.009)
Extreme opinion, high variation	-0.618^{***}	-0.573^{***}
	(0.008)	(0.016)
Extreme opinion, low variation	-0.775^{***}	-1.180^{***}
	(0.021)	(0.044)
Moderate opinion, low variation	-0.206^{***}	-0.309^{***}
	(0.006)	(0.012)
Neighborhood size (scaled)	-0.004^{***}	-0.040^{***}
	(0.0005)	(0.001)
Elite Positive	0.021^{***}	0.040^{***}
	(0.001)	(0.003)
Elite Negative	0.004^{***}	0.020^{***}
	(0.001)	(0.002)
Interaction Ratio × Ext. high var.	0.198^{***}	0.572^{***}
	(0.009)	(0.020)
Interaction Ratio × Ext. low var.	0.077^{***}	-0.037
	(0.023)	(0.047)
Interaction Ratio × Mod. low var.	0.240^{***}	0.182^{***}
	(0.005)	(0.011)
Constant	0.610^{***}	0.874^{***}
	(0.005)	(0.011)
Observations	1,800	1,800
R^2	0.990	0.892
Adjusted R^2	0.989	0.891
Residual Std. Error (df = 1789)	0.016	0.035
F Statistic (df = 10; 1789)	$16{,}951.050^{***}$	$1{,}469.959^{***}$
Note:		*p<0.05; **p<0.01; ***p<0.001

Table 8: Regression Results: Camp-Switching Rate and Mean Absolute Change (Willingness, 20%)

	Dependent variable:	
	Camp-Switching Rate (CSR) CSR	Mean Absolute Change (MAC) MAC
Interaction Ratio (scaled)	-0.148^{***}	-0.031^{***}
	(0.008)	(0.007)
Extreme opinion, high variation	-0.346^{***}	1.240^{***}
	(0.012)	(0.011)
Extreme opinion, low variation	-0.641^{***}	0.243^{***}
	(0.034)	(0.031)
Moderate opinion, low variation	0.108^{***}	0.046^{***}
	(0.009)	(0.008)
Neighborhood size (scaled)	-0.006^{***}	-0.022^{***}
	(0.0005)	(0.0004)
Elite Positive	0.029^{***}	0.035^{***}
	(0.001)	(0.001)
Elite Negative	0.007^{***}	0.013^{***}
	(0.001)	(0.001)
Interaction Ratio × Ext. high var.	0.089^{***}	1.257^{***}
	(0.013)	(0.012)
Interaction Ratio × Ext. low var.	-0.162^{***}	0.289^{***}
	(0.035)	(0.031)
Interaction Ratio × Mod. low var.	0.020^{*}	-0.045^{***}
	(0.009)	(0.008)
Constant	0.339^{***}	0.112^{***}
	(0.008)	(0.007)
Observations	1,800	1,800
R^2	0.979	0.967
Adjusted R^2	0.979	0.967
Residual Std. Error (df = 1789)	0.018	0.017
F Statistic (df = 10; 1789)	$8{,}206.303^{***}$	$5{,}245.465^{***}$
Note:		*p<0.05; **p<0.01; ***p<0.001

Table 9: Regression Results: Camp-Switching Rate and Mean Absolute Change (Friend-Learning)

	Dependent variable:	
	Camp-Switching Rate (CSR) CSR	Mean Absolute Change (MAC) MAC
Interaction Ratio (scaled)	-0.275^{***}	-0.318^{***}
	(0.004)	(0.006)
Extreme opinion, high variation	-0.606^{***}	0.218^{***}
	(0.007)	(0.010)
Extreme opinion, low variation	-0.822^{***}	-0.708^{***}
	(0.020)	(0.030)
Moderate opinion, low variation	-0.155^{***}	-0.184^{***}
	(0.005)	(0.007)
Neighborhood size (scaled)	-0.006^{***}	-0.031^{***}
	(0.0003)	(0.001)
Elite Positive	0.027^{***}	0.046^{***}
	(0.001)	(0.001)
Elite Negative	0.007^{***}	0.020^{***}
	(0.001)	(0.001)
Friend-Learning (2.5 vs. 1.5)	-0.0001	-0.0002
	(0.0003)	(0.0004)
Interaction Ratio $\times$ Ext. high var.	0.182^{***}	0.875^{***}
	(0.008)	(0.012)
Interaction Ratio $\times$ Ext. low var.	0.006	-0.025
	(0.021)	(0.032)
Interaction Ratio $\times$ Mod. low var.	0.228^{***}	0.138^{***}
	(0.005)	(0.007)
Constant	0.565^{***}	0.516^{***}
	(0.005)	(0.007)
Observations	3,600	3,600
R^2	0.985	0.885
Adjusted R^2	0.985	0.885
Residual Std. Error (df = 3588)	0.018	0.027
F Statistic (df = 11; 3588)	$22{,}144.790^{***}$	$2{,}508.347^{***}$
Note:		*p<0.05; **p<0.01; ***p<0.001

Model diagnostics

A.1.4 The Bayesian models

We evaluated the Bayesian CSR model using standard regression diagnostics (Figure 5). The posterior predictive check indicates that the model reproduces the overall distribution of the CSR well, with predicted densities closely tracking the observed distribution across the main peaks.

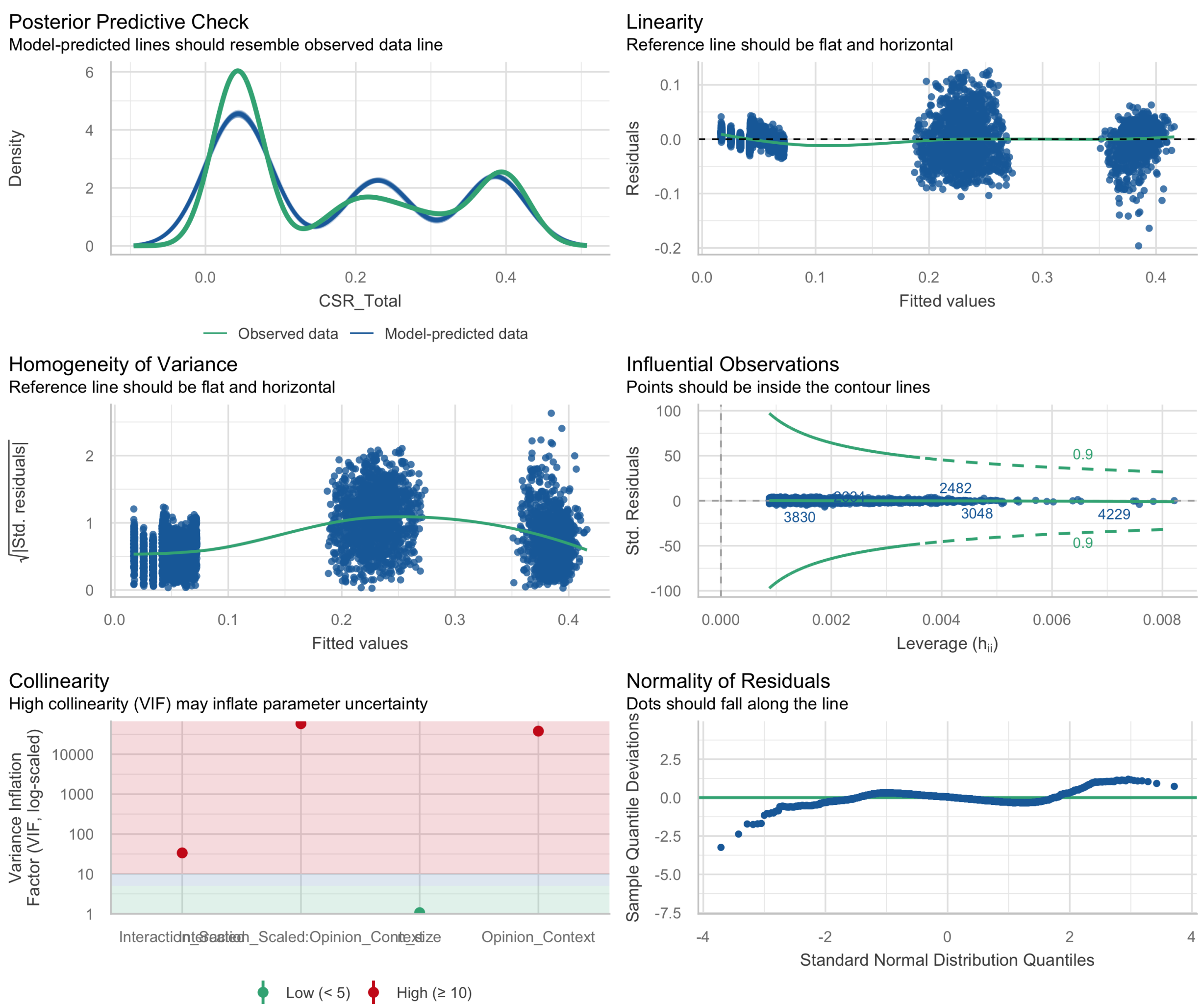

Figure 5: Regression diagnostic plots for the Bayesian CSR model.

Residual–fitted plots show no evidence of systematic nonlinearity; the smoothing line remains approximately flat across the range of fitted values. The scale–location plot suggests mild heteroskedasticity, with somewhat greater dispersion at intermediate fitted values, although the variance pattern remains bounded.

Leverage diagnostics indicate that no individual simulation run exerts disproportionate influence on the estimates, as all observations fall within Cook's distance contours and leverage values remain low. Variance inflation factors are elevated for the contextual indicators and their interaction terms, which is expected in models that include categorical interactions. The main predictor of interest, interaction frequency, exhibits low collinearity.

The Q–Q plot shows minor departures from normality in the residual tails, while the central quantiles closely follow the theoretical normal distribution. Given the large sample size ($N = 5{,}400$), these deviations are unlikely to affect coefficient estimates or statistical inference. Overall, the diagnostics indicate that the model provides an adequate statistical representation of the simulated opinion dynamics.

We evaluated the MAC model using standard regression diagnostics (Figure 6). The posterior predictive check indicates that the model reproduces the overall distribution of Mean Absolute Change (MAC) reasonably well. The predicted density closely follows the observed distribution across the central mass of the data, although the model slightly smooths the sharp peak around MAC ≈ 0.10 and modestly underestimates variation in the right tail.

Residual–fitted plots suggest mild nonlinearity at higher fitted values. The smoothing line remains close to horizontal across most of the range but bends upward at the upper end of predicted MAC values, indicating that the linear specification slightly underestimates larger opinion shifts. The scale–location plot shows moderate heteroskedasticity, with residual dispersion increasing somewhat at higher fitted values. However, the variance pattern remains gradual and bounded.

Leverage diagnostics indicate that no individual simulation run exerts disproportionate influence on the estimated coefficients. All observations fall well within the Cook's distance contours and leverage values remain small.

Variance inflation factors are high for the contextual indicators and their interaction terms. This pattern arises mechanically from the inclusion of categorical interactions and does not indi-

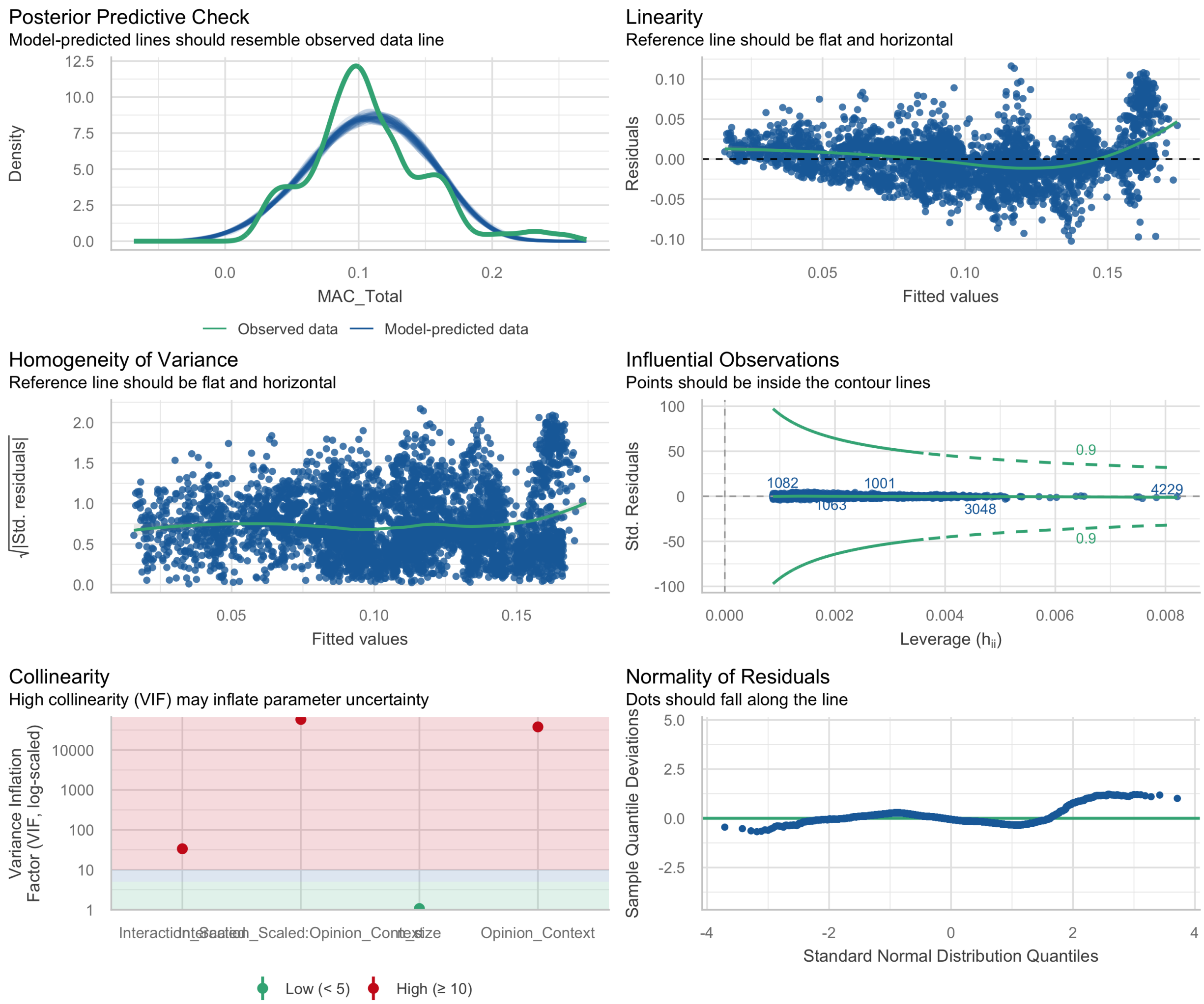

Figure 6: Regression diagnostic plots for the Bayesian MAC model.

cate problematic redundancy among the substantive predictors. The primary explanatory variable, interaction frequency, exhibits low collinearity.

Finally, the Q–Q plot indicates modest departures from normality in the upper tail of the residual distribution. Residuals follow the theoretical normal line closely across the central quantiles, with deviations concentrated among a small number of high-MAC observations. Given the large sample size ($N = 5{,}400$), these deviations are unlikely to materially affect coefficient estimates or inference.

Overall, the diagnostics suggest that the regression specification provides an adequate statistical approximation of the simulated dynamics governing the magnitude of opinion change.

A.1.5 The Backlash models

Figures 7 and 8 report diagnostic tests evaluating the adequacy of the backlash regression models for the Camp Switching Rate (CSR) and Mean Absolute Change (MAC), respectively.

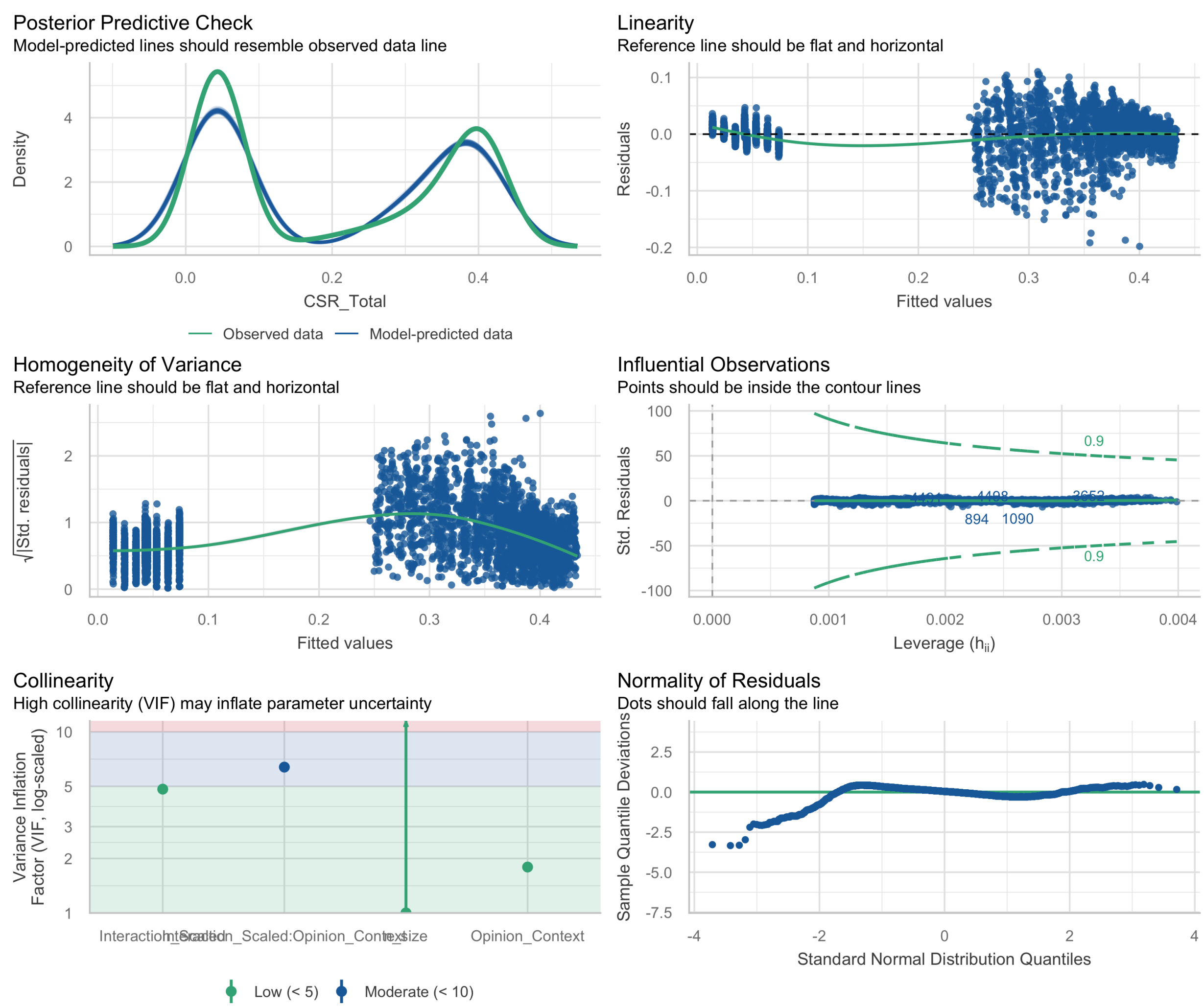

Figure 7: Regression diagnostic plots for the backlash CSR model.

Figure 7 shows the diagnostics for the CSR model. The posterior predictive check indicates that the model reproduces the bimodal structure of the observed CSR distribution reasonably well,

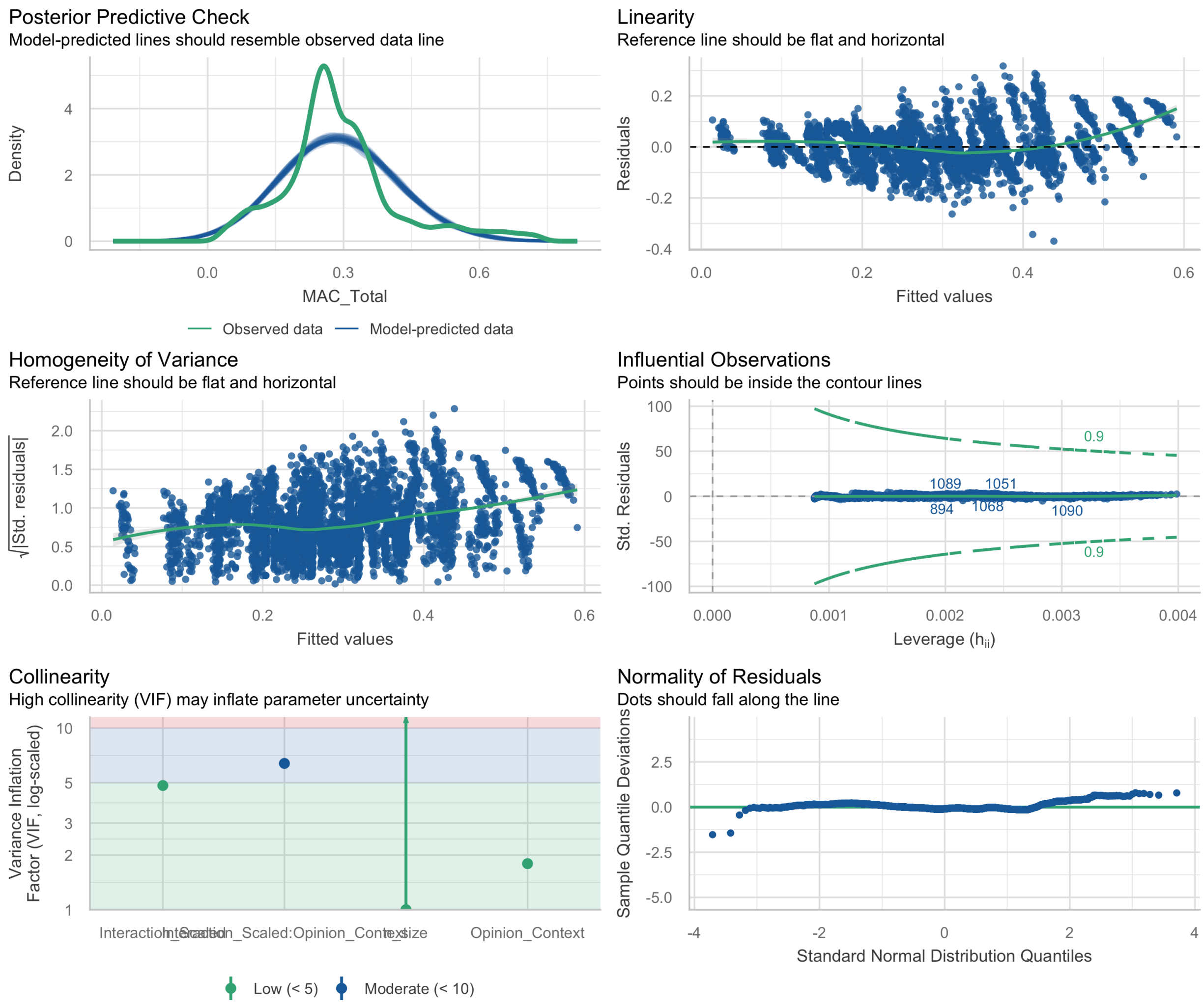

Figure 8: Regression diagnostic plots for the backlash MAC model.

capturing both the low-switching and high-switching regimes present in the simulation output. The residual–fitted plot suggests that the linear specification is broadly appropriate: the smoothing line remains close to horizontal across most fitted values, with only mild curvature at the upper end. The scale–location plot indicates moderate heteroskedasticity, with somewhat greater residual dispersion among higher fitted CSR values, although the variance pattern remains bounded. The leverage diagnostics show that no simulation run exerts disproportionate influence on the estimated coefficients, as all observations lie well within Cook's distance contours and leverage values remain small. Variance inflation factors indicate moderate collinearity among the contextual indicators and their interaction terms, which arises mechanically from the inclusion of categorical interactions. The Q–Q plot shows mild departures from normality in the residual tails, while the central quantiles follow the theoretical normal distribution closely.

Figure 8 presents the diagnostics for the MAC model. The posterior predictive check indicates that the model reproduces the central mass of the observed MAC distribution, although the predicted density slightly smooths the sharp peak present in the data. The residual–fitted plot reveals somewhat stronger curvature than in the CSR model, with the smoothing line bending upward at higher fitted values. This pattern suggests that the linear specification may modestly underestimate larger opinion shifts. The scale–location plot indicates moderate heteroskedasticity, with residual variance increasing as predicted MAC values grow larger. As with the CSR model, leverage diagnostics indicate that no single observation exerts undue influence on the regression estimates. Variance inflation factors again reflect moderate collinearity among the contextual indicators and interaction terms, while the principal explanatory variable remains well identified. The Q–Q plot shows modest departures from normality in the upper tail of the residual distribution, although the central quantiles closely follow the theoretical normal line.

Given the large number of simulation runs ($N = 5{,}400$), these deviations from ideal linear model assumptions are unlikely to materially affect coefficient estimates or statistical inference. The diagnostics therefore indicate that the regression models provide a reasonable statistical representation of the opinion dynamics generated by the backlash mechanism.

A.1.6 The Elite cues models under Bayesian updating

Figures 9 and 10 report regression diagnostics for the elite signaling models predicting the CSR and MAC, respectively. According to Figure 9, the posterior predictive check indicates that the model captures the main features of the observed CSR distribution, including the prominent peak at low switching rates and the secondary cluster at higher switching values. The predicted density slightly smooths the observed peaks but closely tracks the overall shape of the distribution.

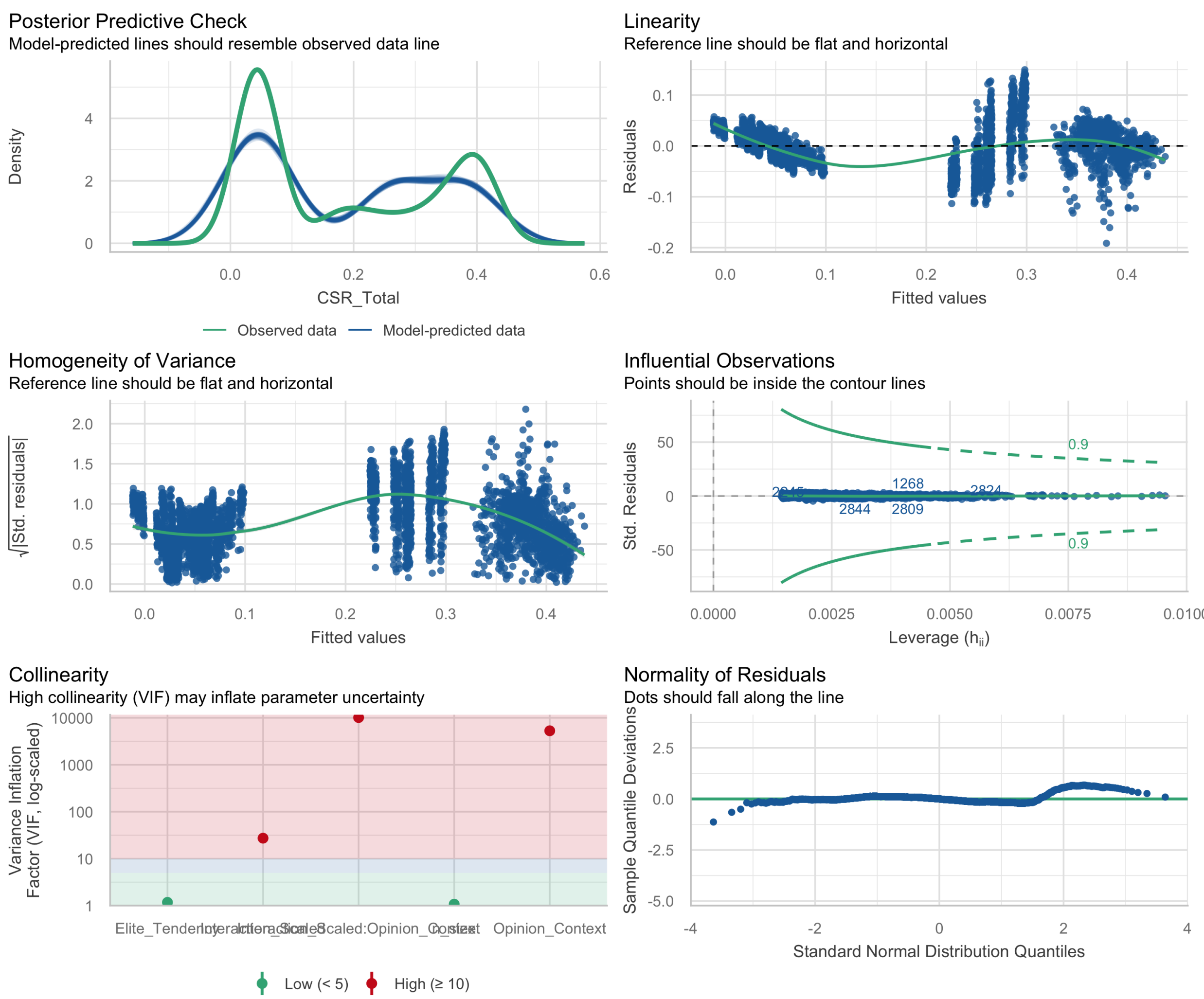

Figure 9: Regression diagnostic plots for the Elite cues CSR model (Bayesian).

The residual–fitted plot suggests that the linear specification provides a reasonable approximation of the relationship between predictors and CSR. The smoothing line remains close to horizontal across most of the fitted values, although a slight downward trend appears at the highest predicted CSR levels. The scale–location plot indicates moderate heteroskedasticity, with residual dispersion increasing somewhat among intermediate CSR predictions before declining at the highest fitted values.

Leverage diagnostics show that no individual simulation run exerts undue influence on the regression estimates. All observations remain well within Cook's distance contours and leverage

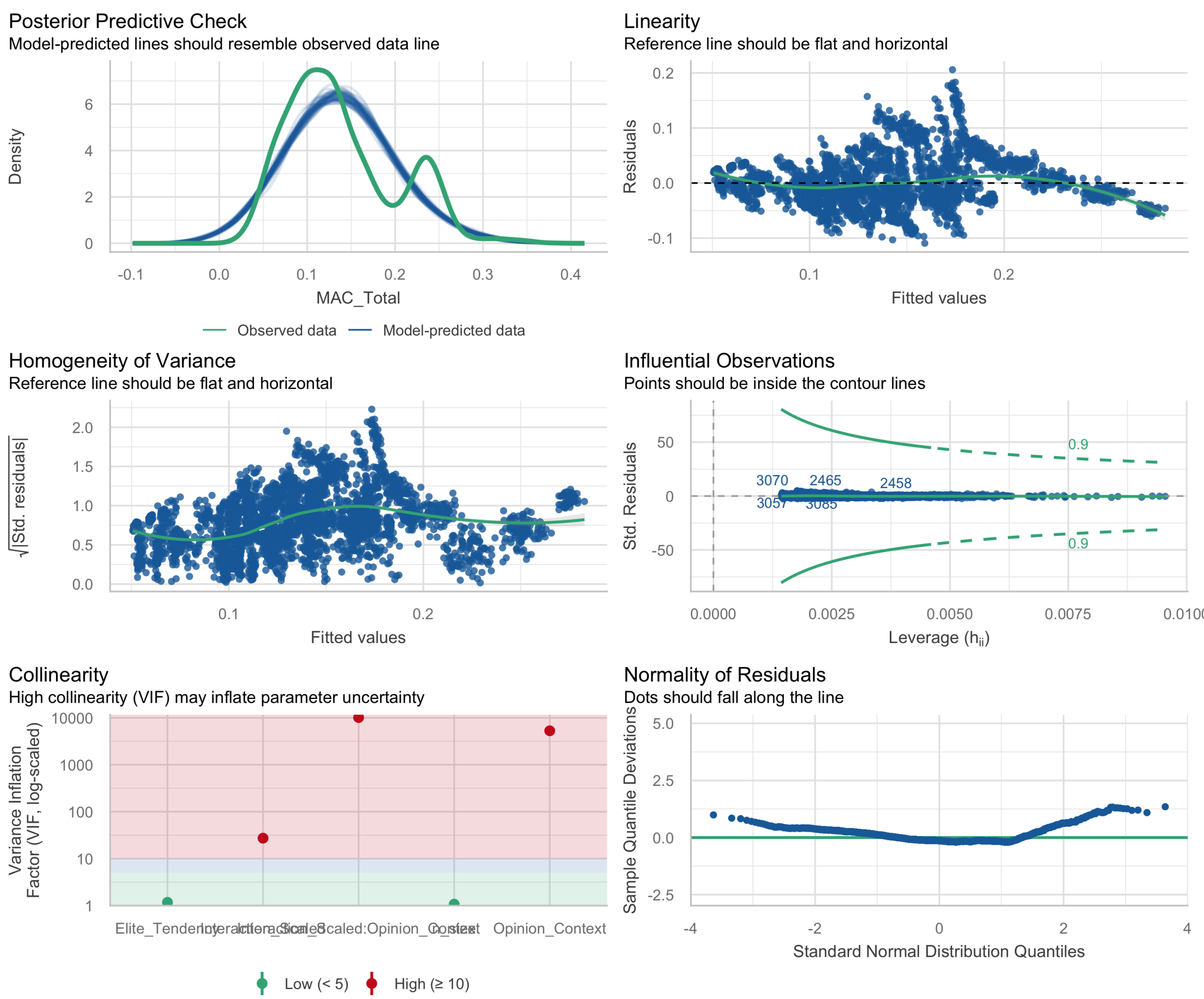

Figure 10: Regression diagnostic plots for the Elite cues MAC model (Bayesian).

values are small. Variance inflation factors indicate elevated collinearity among the contextual indicators and their interaction terms, which arises from the inclusion of categorical interactions rather than from problematic redundancy among the predictors. The Q–Q plot shows modest departures from normality in the residual tails, while the central quantiles align closely with the theoretical normal distribution.

According to Figure 10, the posterior predictive check shows that the model reproduces the overall distribution of opinion change reasonably well, capturing both the central peak and the right-hand tail associated with larger opinion shifts. The predicted distribution slightly smooths the sharp peak present in the observed data.

The residual–fitted plot indicates mild curvature at higher fitted values, suggesting that the linear specification modestly underestimates larger magnitudes of opinion change. The scale–location plot similarly indicates moderate heteroskedasticity, with residual dispersion increasing gradually as predicted MAC values grow larger.

Influence diagnostics again indicate that no single observation exerts disproportionate leverage on the regression estimates. Variance inflation factors reveal substantial collinearity among contextual indicators and their interaction terms, a pattern expected in models including categorical interactions. The principal explanatory variables remain identifiable. The Q–Q plot indicates modest departures from normality in the extreme tails of the residual distribution, while the central quantiles follow the theoretical normal line closely.

Given the large number of simulation runs ($N = 5{,}400$), these deviations from ideal linear model assumptions are unlikely to materially affect coefficient estimates or statistical inference. The diagnostics therefore suggest that the regression specifications provide an adequate statistical representation of the opinion dynamics generated by the elite signaling mechanism.

A.1.7 The Elite cues under the backlash mechanism

Figures 11 and 12 present regression diagnostics for the backlash models that incorporate elite signaling, predicting the CSR and MAC, respectively.

According to Figure 11, the posterior predictive check indicates that the model reproduces the key features of the CSR distribution generated by the simulation. The predicted density closely follows the bimodal structure of the observed data, capturing both the low-switching cluster and the higher switching regime associated with polarized interaction environments.

The residual–fitted plot suggests that the linear specification provides a reasonable approximation of the relationship between the predictors and CSR. The smoothing line remains close to horizontal across most of the fitted values, with only mild curvature in the middle of the dis-

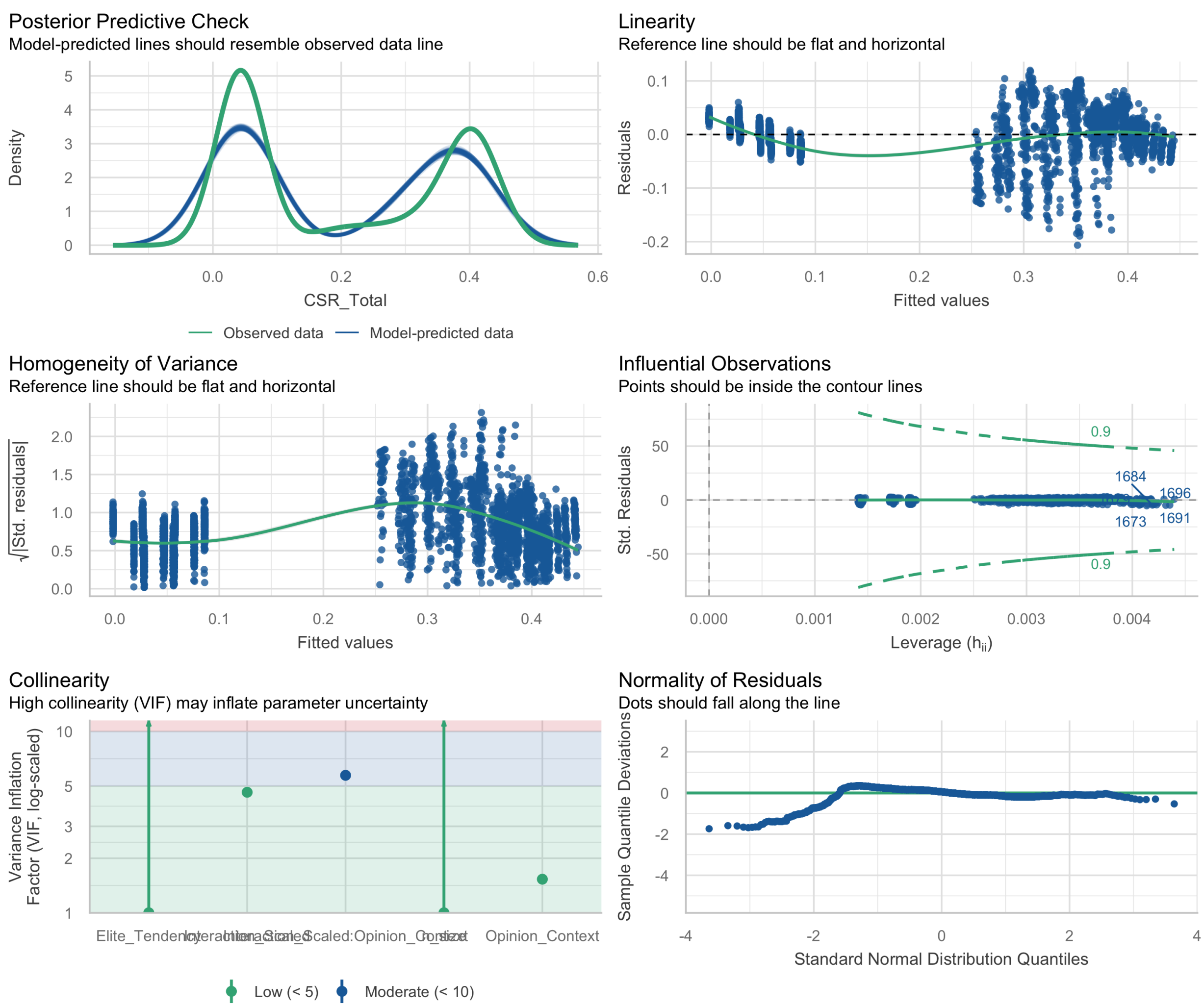

Figure 11: Regression diagnostic plots for the Elite cues CSR model (Backlash).

tribution. The scale–location plot indicates moderate heteroskedasticity, with residual dispersion increasing among intermediate fitted values before declining slightly at the upper end of the CSR range.

Influence diagnostics show that no individual simulation run exerts disproportionate leverage on the estimated coefficients. All observations lie well within Cook's distance contours and leverage values remain small across the sample. Variance inflation factors indicate moderate collinearity among contextual indicators and interaction terms. This pattern arises mechanically from the inclusion of categorical interactions rather than from problematic redundancy among the predictors.

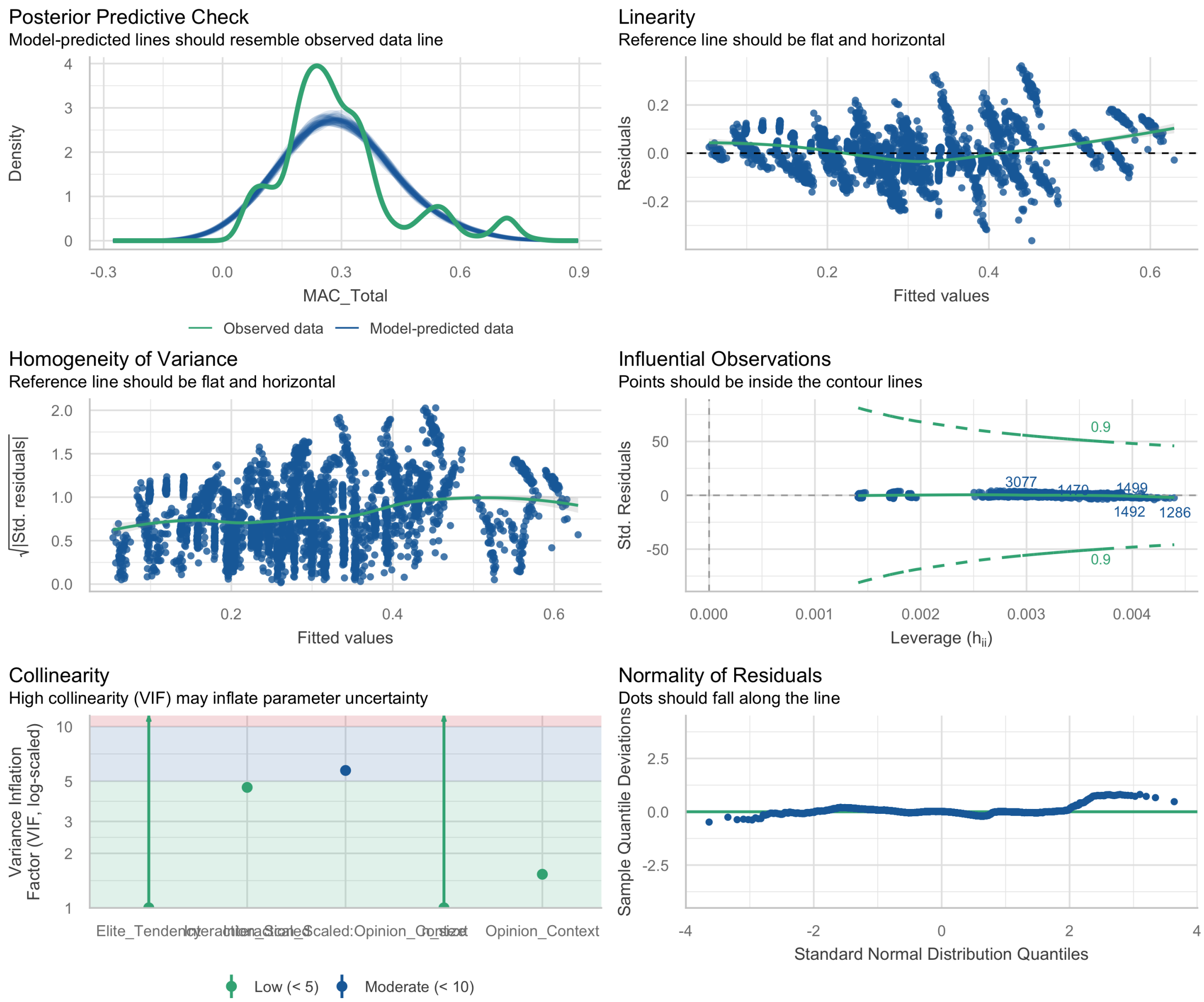

Figure 12: Regression diagnostic plots for the Elite cues MAC model (Backlash).

The Q–Q plot indicates mild departures from normality in the residual tails, while the central quantiles track the theoretical normal distribution closely.

According to Figure 12, the posterior predictive check shows that the model captures the overall distribution of mean opinion change reasonably well, reproducing the central mass of the observed distribution while slightly smoothing the sharp peak generated by the simulation.

The residual–fitted plot indicates modest curvature at higher fitted values, suggesting that the linear specification slightly underestimates the largest opinion shifts. The scale–location plot similarly shows moderate heteroskedasticity, with residual variance gradually increasing as predicted MAC values grow larger.

As with the CSR model, leverage diagnostics indicate that no single observation exerts undue influence on the regression estimates. Variance inflation factors again reflect moderate collinearity among contextual indicators and their interaction terms, which is expected given the interaction structure of the model. The Q–Q plot shows modest deviations from normality in the extreme tails of the residual distribution, while the central portion of the distribution follows the theoretical normal line closely.

Given the large number of simulation runs ($N = 5{,}400$), these deviations from ideal linear model assumptions are unlikely to materially affect coefficient estimates or statistical inference. The diagnostics therefore suggest that the regression specifications provide an adequate statistical representation of the opinion dynamics produced by the backlash mechanism under elite signaling.