Disciplined Discretion: Premise-Based Governance of Expert Overrides in Algorithm-Augmented Work

Sangseok Lee

Bayes Business School, City, University of London, London, UK
sangseok.lee@bayes.city.ac.uk

1 Problem and Research Question

A central problem of algorithm-augmented work is that experts retain discretion over algorithmic advice that, on average, is a strong benchmark. Expert judgment is boundedly rational [1, 2]: it is prone to overconfidence and, in advice settings, to egocentric discounting, the overweighting of one's own opinion relative to advice [6, 7]. The problem is not that experts deviate, since deviations can carry valid local information, but that poorly calibrated, overconfident deviations subtract value [8, 9]. How can an organization govern such discretion? It cannot simply remove it, since local information would be lost, and outcome-based control is weak where outcomes are stochastic and ambiguous, so a single forecast reveals little about judgment quality [3]. The Carnegie tradition offers a third route: organizations shape behavior not by dictating decisions but by designing the premises of decision, that is, what members must attend to and consider [1, 2]. We ask: when an organization installs a justification premise for departing from algorithmic advice, how is the exercise of expert discretion disciplined, and with what performance implications?

2 Theory and Hypotheses

We theorize process accountability as a concrete form of decision-premise control: requiring forecasters to justify how they reasoned, rather than grading them on outcomes, changes what they must be prepared to defend when overriding the model. Process accountability is established to induce more vigilant, self-critical processing, reducing overconfidence and noise unrelated to the target [4, 5]. Applied to discretion over algorithmic advice, it should curb the overconfident component of deviation while leaving informationally grounded deviation intact. We predict that deviation from the algorithm decreases (H1), and, more importantly, that the reduction is *selective*: the share of value-destroying deviations falls while value-creating deviations are preserved (H2). H2 separates premise control from two rivals. Mechanical anchoring predicts uniform compression of deviations toward the model; conformity predicts uniform suppression that also curbs value-creating deviations. Only premise control through justification predicts a selective pattern.

3 Setting, Data, and Identification

Effective 6 October 2023, a Central Region Supplement (CRS 09-2023) required the region's Weather Forecast Offices (WFOs) to document key reasoning, reference the National Blend of Models (NBM) baseline, and quantify uncertainty: a process requirement embedded in the forecast workflow, not an ex-post report. We build a forecast-level panel for 66 WFOs (35 treated, 31 control), 2021 to 2024, linking each office's edited forecast (PFM) to the NBM baseline and to station observations (ASOS). Reliance is adherence to the baseline, $-|\text{PFM}-\text{NBM}|$; performance is override value, $|\varepsilon_{\text{NBM}}|-|\varepsilon_{\text{PFM}}|$, and the share of value-destroying overrides. We estimate $y_{it} = \tau(\text{Treated}_i \times \text{Post}_t) + \alpha_i + \delta_t + \varepsilon_{it}$ with office and month fixed effects. Because treatment is assigned at the region level, we assess inference with wild cluster bootstrap and randomization across regions and placebo dates. To separate the regime from who forecasts, we recover each discussion's authoring forecaster and re-estimate within individuals present before and after.

4 Preliminary Results

The directive disciplined discretion rather than suppressing it. Deviations from the algorithm fell by 1.24°F and forecast accuracy improved (Fig. 1; both $p < .001$). Decomposing the override value shows the gain was selective (Fig. 2): the value destroyed by overrides fell sharply (1.05°F, $p < .001$), while the value added by overrides was preserved and even rose slightly (0.16°F, $p < .01$). A mechanical anchoring or conformity account would compress all overrides, reducing value-creating ones too; instead, only the value-destroying component fell. Mechanism tests reinforce this: the policy's measurable behavioral footprint was the justification premise, as structured reasoning rose sharply, whereas algorithm referencing did not change and does not track the outcome. Finally, the effect is essentially identical within individual forecasters (706 of 1,191 span the policy), ruling out a compositional artifact.

5 Contribution

We make three contributions. First, we extend Carnegie's decision-premise control to algorithm-augmented expert work: where organizations can neither remove discretion nor rely on outcome control, they govern judgment by shaping the premises under which deviations from algorithmic advice must be justified. Second, premise control disciplines discretion selectively rather than suppressing it uniformly. Third, reliance on algorithmic advice depends not only on individual trust, expertise, or model performance, but also on the organizational premises embedded in the decision process.

References

1. Simon, H.A.: Administrative Behavior. Macmillan, New York (1947)
2. March, J.G., Simon, H.A.: Organizations. Wiley, New York (1958)
3. Thompson, J.D.: Organizations in Action. McGraw-Hill, New York (1967)
4. Siegel-Jacobs, K., Yates, J.F.: Effects of procedural and outcome accountability on judgment quality. Organ. Behav. Hum. Decis. Process. **65**(1), 1–17 (1996)
5. Tetlock, P.E.: Accountability theory. In: Thompson, L.L., Levine, J.M., Messick, D.M. (eds.) Shared Cognition in Organizations, pp. 117–137. Erlbaum, Mahwah (1999)
6. Yaniv, I.: Receiving other people's advice. Organ. Behav. Hum. Decis. Process. **93**(1), 1–13 (2004)
7. Bonaccio, S., Dalal, R.S.: Advice taking and decision-making. Organ. Behav. Hum. Decis. Process. **101**(2), 127–151 (2006)
8. Dietvorst, B.J., Simmons, J.P., Massey, C.: Algorithm aversion. J. Exp. Psychol. Gen. **144**(1), 114–126 (2015)
9. Logg, J.M., Minson, J.A., Moore, D.A.: Algorithm appreciation. Organ. Behav. Hum. Decis. Process. **151**, 90–103 (2019)

Appendix

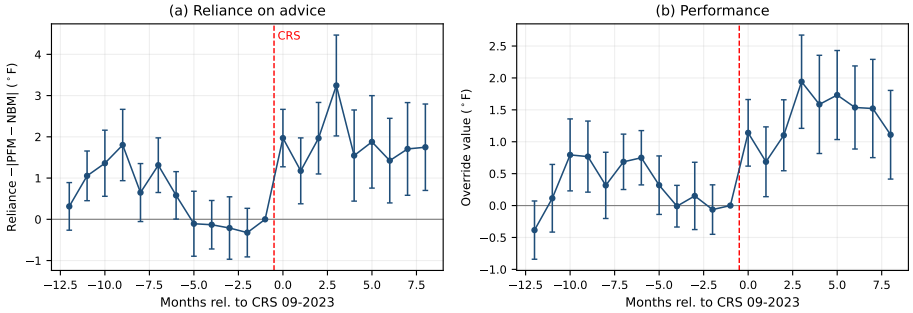

Fig. 1. Event-study estimates by month relative to CRS 09-2023, with 95% confidence intervals (office and month fixed effects, clustered by office). (a) reliance, $-|\mathrm{PFM}-\mathrm{NBM}|$; (b) performance, override value. Pre-policy coefficients show no trend anticipating the post-policy shift.

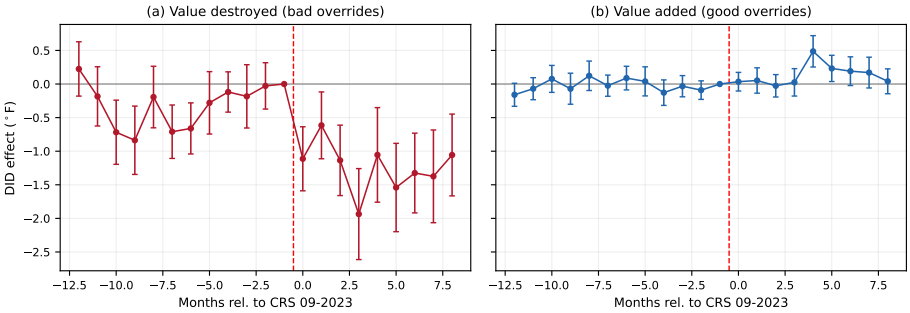

Fig. 2. Selectivity of the effect. Override value is split into two non-negative components and each is estimated as an event study: (a) value destroyed by overrides, $\max(0, |\varepsilon_{\mathrm{PFM}}| - |\varepsilon_{\mathrm{NBM}}|)$; (b) value added by overrides, $\max(0, |\varepsilon_{\mathrm{NBM}}| - |\varepsilon_{\mathrm{PFM}}|)$. The policy sharply reduces value destroyed while leaving value added intact, indicating selective rather than uniform discipline of discretion.