The Reputational Tax of AI: How Structural Support and Incentives Shape Employee Disclosure Behavior

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

As enterprises integrate artificial intelligence, the fidelity of productivity metrics is threatened by employees' strategic misreporting of their AI usage. This behavior arises from a "reputational tax" associated with algorithmic uncertainty, compelling employees to either conceal AI use to avoid blame for errors or performatively overstate it to signal technological fluency. To dissect the drivers of this behavior, we model the choice between "Concealment," "Performative" disclosure, and transparent reporting using multinomial logistic regression on survey data from 2,395 active AI users. The analysis reveals that while perceived AI error frequency drives both forms of misreporting, their underlying motivations are distinct: concealment is a defensive reaction to job insecurity, a pressure exacerbated by the deployment of autonomous agentic systems, whereas performative disclosure is an opportunistic strategy fueled by intrinsic rewards like peer recognition. Crucially, our model demonstrates that concrete structural support—clear AI strategies, training, and safeguards—is a powerful mitigator of both misreporting behaviors, proving substantially more effective than abstract cultural initiatives like promoting learning safety. These findings indicate that to achieve reliable measurement of AI's impact, organizations must prioritize the implementation of robust policy and structural frameworks over purely cultural interventions.

1 Introduction

The integration of artificial intelligence into enterprise workflows is rapidly reshaping the landscape of modern work, promising significant gains in productivity and innovation. As organizations allocate substantial resources to these technologies, the ability to accurately measure their return on investment becomes paramount. However, the very data needed to assess AI's impact is often mediated by the employees who use it, introducing a critical layer of human behavior that can distort measurement. The fidelity of enterprise AI metrics is therefore contingent not only on the technology itself, but on understanding

and managing the strategic responses of the workforce. This study addresses a central challenge in this new environment: the emergence of strategic misreporting, where employees either conceal or exaggerate their use of AI, thereby corrupting the data required for sound organizational decision-making.

We propose that the uncertainty inherent in current AI systems imposes a "reputational tax" on employees, creating a behavioral dilemma. When AI tools are perceived as fallible or their outputs are difficult to trust, employees must strategically manage their association with the technology to protect their professional standing. This can lead to two distinct, yet related, forms of misreporting. On one hand, an employee might engage in concealment, downplaying their reliance on AI to avoid being blamed for potential algorithmic errors. On the other hand, an employee might engage in performative disclosure, overstating their AI usage to signal technological competence and align with organizational pressures to innovate. Both behaviors, while opposite in their expression, stem from the same root cause—algorithmic uncertainty—and both systematically undermine an organization's ability to gauge the true value of its AI investments.

To move beyond anecdotal evidence, this research quantitatively models the determinants of these disclosure behaviors. Using survey data from 2,395 active AI users, we employ a multinomial logistic regression to analyze the choice between transparent reporting, concealment, and performative disclosure. We investigate how this decision is shaped by factors such as the perceived frequency of AI errors, an employee's sense of job security, and the deployment of more autonomous, agentic AI systems. A central aim of our analysis is to disentangle the effectiveness of different organizational interventions, specifically comparing the impact of concrete structural supports—such as clear AI strategies, training programs, and safeguards—with that of more abstract cultural initiatives designed to foster psychological safety and open communication.

Our findings reveal that while the perceived frequency of AI errors is a common catalyst for both forms of misreporting, their underlying motivations are fundamentally different. Concealment emerges as a defensive strategy, strongly predicted by job insecurity and exacerbated by the presence of agentic AI. In contrast, performative disclosure is an opportunistic behavior, primarily driven by the pursuit of intrinsic rewards like peer recognition. Critically, our model demonstrates that robust structural frameworks are the most powerful tool for mitigating both types of misreporting, proving substantially more effective than cultural interventions alone. By identifying the specific organizational levers that foster transparency, this research offers an evidence-based guide for leaders seeking to build an environment where the impact of artificial intelligence can be measured accurately and managed effectively.

2 Methods

2.1 Dataset and sample definition

The data for this study were drawn from a survey of enterprise employees. Our analytical sample was restricted to "Active AI Users," defined as respondents who reported using AI tools at their workplace at a frequency of "Once a week" or higher. This filtering process yielded a final sample of 2,395 respondents for the primary analysis. A comparative group of 208 "Low-Frequency Users" was used to assess potential selection bias. While independent t-tests and Chi-square tests revealed significant differences in demographic profiles such as age and job level between the active and low-frequency groups, no significant differences were found for organizational-level variables, including company size and annual revenue. This finding mitigates concerns that our results are confounded by organizational scale, allowing us to proceed without complex selection bias corrections.

2.2 Variable construction and measurement

The primary analytical model relies on a set of variables constructed from the survey data to capture employee behaviors, organizational context, and individual perceptions.

2.2.1 Dependent variable

The dependent variable for our analysis is a categorical measure of AI disclosure behavior, with three mutually exclusive states:

- **Transparent Reporting (0):** The baseline category, representing employees who do not engage in misreporting their AI usage.
- **Concealment (1):** Representing employees who report downplaying or hiding their use of AI.
- **Performative Disclosure (2):** Representing employees who report overstating or exaggerating their use of AI.

2.2.2 Key independent variables

Our primary predictors were constructed as follows:

- **Foundational Support:** This structural index measures the extent to which an organization provides clear AI strategies, training, and safeguards. As an initial Confirmatory Factor Analysis (CFA) did not converge, we used Principal Component Analysis (PCA) to create the index. The first principal component, which explained 31.12% of the variance, was used. The resulting index demonstrated strong internal consistency (Cronbach's $\alpha = .797$).

- **Cultural Indices:** To distinguish structural support from cultural environment, two additional indices were created via PCA: "Learning Safety," which captures the perceived safety in experimenting and making mistakes, and "Candid Communication," which measures norms around open and honest dialogue regarding workplace challenges.
- **Perceived AI Error Frequency:** A numerical variable capturing the respondent's subjective assessment of how often AI tools produce erroneous or unreliable outputs.
- **Job Security Confidence:** A categorical variable measuring an employee's confidence that their job is secure from being replaced by AI.
- **Agentic AI Deployment:** A binary indicator signifying whether the respondent's organization has deployed autonomous, agentic AI systems.
- **Organizational Incentives:** Two binary indicators were created to capture the presence of "Extrinsic Rewards" (e.g., financial bonuses, career progression) and "Intrinsic Rewards" (e.g., peer recognition, learning opportunities) for AI usage.

2.3 Analytical strategy

To model the determinants of an employee's choice between Transparent, Concealment, and Performative disclosure behaviors, we employed a Multinomial Logistic Regression. This approach is well-suited for analyzing categorical dependent variables with more than two unordered outcomes. The model estimates the log-odds of being in one category of the dependent variable versus the reference category (Transparent Reporting). The general form of the model for outcome j relative to the reference outcome is:

$$\ln\left(\frac{P(Y_i = j)}{P(Y_i = \text{Transparent})}\right) = \beta_{j0} + \beta_{j1}X_{i1} + \cdots + \beta_{jk}X_{ik} \tag{1}$$

where Y_i is the disclosure behavior for individual i, and X_{ik} are the independent variables.

Our analysis proceeded in stages. First, we estimated a base model including perceived error frequency, job security, and the Foundational Support index. Second, to test the "Agentic Shift" hypothesis, we introduced an interaction term between Agentic AI Deployment and Perceived AI Error Frequency. Third, we incorporated the cultural indices (Learning Safety and Candid Communication) to assess their explanatory power relative to structural support. Finally, we conducted a sensitivity analysis on the Performative disclosure outcome to disentangle the effects of extrinsic versus intrinsic rewards.

2.4 Evaluation and model diagnostics

The overall fit of our models was assessed using the Log-Likelihood, Pseudo R^2 values, and Likelihood Ratio tests to compare nested models. To ensure

the robustness of our findings, we performed several diagnostic checks. First, we calculated Variance Inflation Factors (VIFs) for all predictors to test for multicollinearity; all VIFs were well below the conventional threshold of 5, indicating no significant collinearity issues. Second, we tested the Independence of Irrelevant Alternatives (IIA) assumption, a key requirement for multinomial logit models, using a Hausman-McFadden test. The test was non-significant, confirming the suitability of our chosen model specification. Finally, to directly compare the relative importance of structural versus cultural factors, we estimated a fully standardized model, which allowed for a comparison of the effect sizes of the predictors on a common scale.

3 Results

3.1 Sample characteristics and selection bias

Our analysis began by segmenting the survey respondents into "Active AI Users" and "Low-Frequency Users." Active users, defined as those employing AI tools at least "Once a week," constituted the primary analytical sample of 2,395 individuals. The remaining 208 respondents formed the Low-Frequency comparison group.

To assess potential selection bias, we conducted independent Welch's t-tests and Chi-square tests of independence across demographic and organizational variables. The results indicated significant demographic differences between the groups. Active users were, on average, younger ($M = 36.19$, $SD = 9.28$) than Low-Frequency users ($M = 42.12$, $SD = 12.73$), a statistically significant difference, $t(226.51) = -6.56$, $p < .001$. This age gap corresponded to a difference in professional experience, with Active users having fewer years in the workforce ($M = 10.96$, $SD = 8.79$) compared to their counterparts ($M = 18.41$, $SD = 12.89$), $t(224.04) = -8.17$, $p < .001$.

Significant differences were also observed in occupational roles ($\chi^2(10) = 71.07$, $p < .001$) and industry sectors ($\chi^2(12) = 90.45$, $p < .001$). The Active user group had a higher concentration of Analysts (21.5%) and non-managerial Executives (17.7%), while the Low-Frequency group was more heavily composed of Individual Contributors (25.5%) and Entry-level staff (23.6%).

Critically for our analysis, no significant differences were found for key organizational scale metrics, including Global Employee Size ($\chi^2(5) = 4.31$, $p = .505$) and Global Annual Revenue ($\chi^2(4) = 6.48$, $p = .166$). This finding suggests that the level of AI adoption is more closely tied to individual and role-specific factors than to the overall size or revenue of the organization. This mitigates concerns that our subsequent models of disclosure behavior are confounded by organizational scale.

3.2 Construction and validation of key predictors

To measure the degree of structural support for AI within an organization, we constructed a "Foundational Support" index. This index was derived from survey items measuring the provision of clear AI strategies, regular training, and well-defined safeguards. As an initial Confirmatory Factor Analysis (CFA) failed to converge on a single latent variable, we employed Principal Component Analysis (PCA). The first principal component, which explained 31.12% of the total variance, was used to create the index. The resulting scale demonstrated strong internal consistency (Cronbach's $\alpha = .797$).

The Foundational Support index was approximately normally distributed ($M = 0.0$, $SD = 1.933$). To validate that this index captures deliberate organizational policy rather than being a simple proxy for resource availability, we regressed it against ordinal measures of company size and revenue. The model explained a negligible amount of variance ($R^2 = .0025$, $F(2, 2392) = 3.00$, $p = .050$), with neither employee size ($\beta = 0.057$, $p = .075$) nor annual revenue ($\beta = 0.004$, $p = .921$) emerging as significant predictors. This confirms that our Foundational Support measure represents a distinct structural construct related to organizational readiness for AI, independent of corporate scale.

3.3 Determinants of AI disclosure behavior

To model the factors influencing an employee's decision to misreport their AI usage, we estimated a multinomial logistic regression. The model predicts the choice between three behavioral states: Transparent reporting (the reference category), Concealment (downplaying AI use), and Performative disclosure (overstating AI use). The overall model was statistically significant and provided a reasonable fit to the data (Log-Likelihood = -1406.6, Pseudo $R^2 = .078$, LLR $\chi^2(10) = 238.6$, $p < .001$).

The results, visualized in Figure 1, provide strong evidence for a "reputational tax" imposed by algorithmic uncertainty. The perceived frequency of AI errors was a significant positive predictor for both forms of misreporting. For each one-unit increase in perceived error frequency, the log-odds of engaging in Concealment increased by 0.160 ($p = .036$), and the log-odds of engaging in Performative disclosure increased by 0.391 ($p < .001$). This dual finding suggests that as employees perceive AI to be less reliable, they strategically adjust their reported usage to manage reputational risk, either by distancing themselves from potential errors (Concealment) or by performatively signaling competence despite the tool's flaws (Performative).

Conversely, the presence of strong Foundational Support acted as a powerful mitigating factor. Higher levels of structural support significantly reduced the log-odds of both Concealment ($\beta = -0.310$, $p < .001$) and Performative disclosure ($\beta = -0.285$, $p < .001$). This indicates that clear strategies, training, and safeguards can effectively lower the reputational tax and encourage more transparent reporting.

An employee's confidence in their job security also played a crucial, albeit

differential, role. Compared to the baseline, employees who were "Somewhat unconfident" ($\beta = 1.005$, $p < .001$) or "Neither/Unsure" ($\beta = 0.908$, $p < .001$) about their job security had significantly higher log-odds of concealing their AI use. This supports the interpretation of Concealment as a defensive, risk-averse strategy. In contrast, being "Very confident" in one's job security significantly reduced the likelihood of Performative disclosure ($\beta = -0.669$, $p < .001$), suggesting that employees who feel secure have less need to engage in opportunistic signaling.

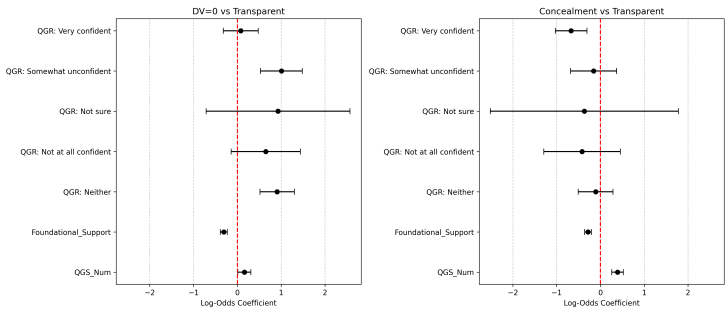

Figure 1: Coefficient plot from the multinomial logistic regression model showing the predictors of AI disclosure behaviors. The panels display the log-odds of engaging in Performative disclosure (left) and Concealment (right) relative to the Transparent reference category, with error bars indicating 95% confidence intervals. The results visualize that increased perceived AI error frequency (QGS_Num) is associated with a higher likelihood of both misreporting behaviors, whereas greater Foundational_Support acts as a significant mitigating factor. Confidence in job security (QGR) shows a differential effect, where lower confidence levels significantly increase the odds of Concealment, while being 'Very confident' reduces the odds of Performative disclosure.

3.4 The moderating role of agentic AI

We next investigated whether the reputational calculus changes with the deployment of more autonomous, agentic AI systems. We tested this "Agentic Shift" hypothesis by introducing an interaction term between the deployment of agentic AI and the perceived frequency of AI errors into our model. The interaction model provided a significantly better fit than the main effects model (Log-Likelihood = -1394.9, Pseudo $R^2 = .086$, LLR $\chi^2(2) = 23.4$, $p < .001$).

The results revealed a critical divergence in behavioral responses. For the Concealment outcome, the interaction between error frequency and the presence of agentic AI was positive and statistically significant ($\beta = 0.384$, $p = .028$). As illustrated in Figure 2, the probability of concealment rises sharply with perceived error frequency, but only in organizations where agentic AI has been deployed. This suggests that the reputational tax is amplified when AI systems

operate with greater autonomy; employees may feel a stronger need to hide their reliance on these systems when they are prone to making errors that cannot be easily intercepted or corrected.

In contrast, this interaction effect was not significant for Performative disclosure ($\beta = 0.013$, $p = .929$). The propensity to overstate AI use in the face of errors does not appear to be exacerbated by the presence of agentic systems. This finding underscores that Concealment and Performative disclosure, while both linked to AI fallibility, are driven by distinct psychological and contextual pressures.

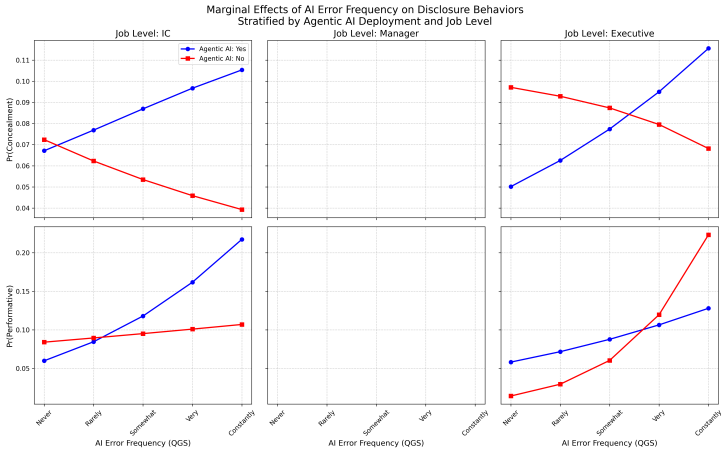

Figure 2: Marginal effects of AI error frequency on the probability of Concealment (top) and Performative (bottom) disclosure, conditioned on the deployment of agentic AI and stratified by job level. The probability of Concealment rises sharply with increasing error frequency when agentic AI is deployed (blue line), an interaction not observed for Performative disclosure. This finding suggests that the reputational tax of AI unreliability is exacerbated for concealment behaviors when systems operate autonomously.

3.5 The paradoxical effect of organizational incentives

To better understand the motivations behind Performative disclosure, we conducted a sensitivity analysis on the role of organizational incentives. We modeled the likelihood of this behavior as a function of whether the organization offers Extrinsic Rewards (e.g., financial bonuses, career progression) or Intrinsic Rewards (e.g., peer recognition, learning opportunities) for AI usage.

The results uncovered a striking paradox. The presence of Extrinsic Rewards had no statistically significant effect on the likelihood of engaging in Performative disclosure ($\beta = 0.026$, $p = .856$). However, the availability of Intrinsic Rewards was a powerful and highly significant predictor, massively increasing the log-odds of this behavior ($\beta = 0.996$, $p < .001$). A likelihood ratio test

confirmed that a model including Intrinsic Rewards provided a vastly superior fit compared to one with only Extrinsic Rewards ($\chi^2(2) = 48.62$, $p < .001$).

This finding suggests that attempts to foster AI adoption through social capital and peer recognition may inadvertently create incentives for "innovation theater." Employees appear to game the system by overstating their AI use to capture these non-monetary, social rewards. In contrast, financial incentives, which are more likely to be tied to verifiable outputs, do not provoke the same performative behavior.

3.6 Comparing the efficacy of structural and cultural support

A central goal of this study was to disentangle the relative effectiveness of concrete structural supports versus abstract cultural initiatives in promoting transparent reporting. To this end, we introduced two cultural indices into our model: "Learning Safety" (capturing the perceived safety to experiment and fail) and "Candid Communication" (measuring norms around open dialogue). While the inclusion of these cultural variables improved the overall model fit (Likelihood Ratio $\chi^2(4) = 27.48$, $p < .001$), their individual effects were divergent.

A culture of Candid Communication significantly deterred both Concealment ($\beta = -0.287$, $p < .001$) and Performative disclosure ($\beta = -0.128$, $p = .045$). However, Learning Safety—a factor often highlighted as crucial for innovation—had no significant effect on reducing either Concealment ($\beta = 0.013$, $p = .849$) or Performative disclosure ($\beta = 0.115$, $p = .111$).

To directly compare the magnitude of structural versus cultural effects, we estimated a fully standardized model focusing on Foundational Support and Learning Safety. The results, shown in Figure 3, unequivocally demonstrate the primacy of structural interventions. For the Concealment outcome, the standardized coefficient for Foundational Support ($\beta = -0.296$) was substantially larger in magnitude than the non-significant effect of Learning Safety ($\beta = 0.078$). This gap was even more pronounced for Performative disclosure, where Foundational Support exerted a strong negative effect ($\beta = -0.460$), while Learning Safety again had no significant impact ($\beta = 0.093$).

These findings collectively argue that while a culture of open dialogue is beneficial, it is the implementation of concrete, structural enablers—clear guidelines, robust training, and defined safeguards—that is most effective in mitigating the reputational risks associated with AI and fostering an environment of transparent reporting.

3.6.1 Model diagnostics and robustness checks

Finally, all models were subjected to diagnostic tests to ensure robustness. Variance Inflation Factors (VIFs) for all predictors were well below the conventional threshold of 5 (maximum VIF = 1.54), indicating no issues with multicollinearity. A Hausman-McFadden test confirmed that the Independence of Irrelevant Alternatives (IIA) assumption of the multinomial logit model was not violated

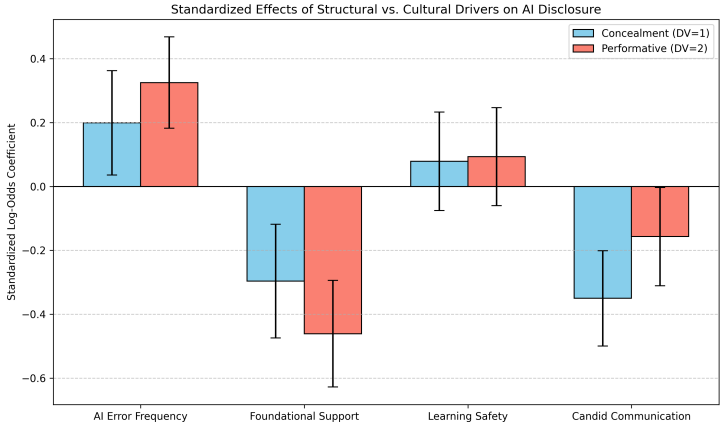

Figure 3: Standardized log-odds coefficients from a multinomial logistic regression model comparing the relative impact of Foundational Support (structural) and Learning Safety (cultural) on AI disclosure behaviors. The plot visually demonstrates that structural interventions, specifically Foundational Support, significantly reduce the likelihood of both Concealment and Performative disclosure. In contrast, the cultural factor of Learning Safety has a non-significant effect on either misreporting behavior. The results highlight the superior efficacy of concrete structural enablers over abstract cultural initiatives in mitigating strategic misreporting of AI use. Error bars represent 95% confidence intervals.

(HM Stat $= 0.0$, $p = 1.0$). Furthermore, sensitivity checks stratifying the sample by high versus low AI usage frequency demonstrated that the core dynamics remained consistent across usage intensities.

4 Conclusions

This study addressed the challenge of accurately measuring the impact of artificial intelligence in the enterprise, a task complicated by employees' strategic misreporting of their AI usage. We proposed that a "reputational tax" arising from algorithmic uncertainty compels employees to either conceal their AI use to avoid blame for errors or performatively overstate it to signal technological fluency. To investigate the drivers of these behaviors, we analyzed survey data from 2,395 active AI users, employing a multinomial logistic regression to model the choice between transparent reporting, concealment, and performative disclosure.

Our analysis revealed that while the perceived frequency of AI errors is a common catalyst for both forms of misreporting, their underlying motivations are distinct. Concealment was found to be a defensive strategy, strongly predicted by an employee's sense of job insecurity. This behavior is significantly exacerbated in organizations that have deployed more autonomous, agentic AI

systems, suggesting that the reputational risk is amplified when AI operates with less human oversight. In contrast, performative disclosure emerged as an opportunistic strategy. We found it was not driven by extrinsic financial incentives but was powerfully predicted by the availability of intrinsic rewards, such as peer recognition, indicating that social pressures to innovate can inadvertently encourage disingenuous reporting.

Crucially, this research provides clear guidance on the effectiveness of different organizational interventions. Our models demonstrate that the implementation of concrete structural support—encompassing clear AI strategies, comprehensive training, and well-defined safeguards—is the most powerful mitigator of both concealment and performative disclosure. While a culture of candid communication also proved beneficial, the widely advocated principle of learning safety had no significant effect on reducing misreporting. A comparison of standardized effect sizes confirmed that structural frameworks are substantially more effective than these abstract cultural initiatives.

In conclusion, this study demonstrates that to obtain a reliable measure of AI's return on investment, organizations must look beyond cultural programs and focus on building robust structural and policy frameworks. By providing clear guidance, training, and safeguards, leaders can effectively lower the reputational tax associated with AI use, thereby fostering an environment where employees report their interactions with these technologies transparently and accurately.