Between Audit and Self-Reference: Content Analysis of Language Models' Self-Assessments of a User–AI Interaction Governance Protocol in Management Research

ABSTRACT

This study investigates how different families of large language models (LLMs) interpret and self-assess a user–AI interaction governance protocol, called the User–GenAI Pact, designed to regulate the use of generative artificial intelligence by master's and doctoral students in Management. The objective was to identify and compare, through Content Analysis, the meanings that the systems themselves attribute to their capabilities, limitations, and the risks of applying the protocol, as well as the representations they construct of one another. A qualitative approach grounded in Bardin was adopted, using a corpus of twelve technical-critical assessments produced by six foundation models (Gemini, Qwen, DeepSeek, Llama, ChatGPT, and Claude) in response to two structured prompts and subjected to a priori and a posteriori categorical coding. The results revealed three emergent categories: the contested stratification of declared technical capabilities; the asymmetry of costs and benefits across the research cycle, favoring theoretical-methodological structuring and disfavoring empirical data collection and analysis; and a systematic pattern of self-referential bias, in which each model tends to attribute greater critical rigor and less complacency to itself and to architecturally similar systems than to its competitors. Limitations include the absence of human intercoder verification, the declarative rather than behavioral nature of the corpus, and its temporal delimitation to August 2026. The study contributes theoretically to the literature on algorithmic governance and organizational trust in intelligent systems, and practically to the formulation of institutional protocols for the use of generative AI in graduate programs.

Keywords: algorithmic governance; generative artificial intelligence; content analysis; organizational trust; graduate education in management.

1. INTRODUCTION

The diffusion of large language models (LLMs) in academic research environments has rapidly altered the routines of knowledge production in Management, shifting part of the tasks of theoretical structuring, writing, and revision to generative artificial intelligence (GenAI) systems operated in conversational mode. In the absence of consolidated institutional norms for this use, a

phenomenon little explored in the management literature until now has emerged: users themselves are devising informal governance protocols—here termed "meta-prompts" or "pacts"—that seek to discipline the interaction between researcher and system turn by turn, specifying rules for classifying requests, confirmation triggers for high-risk tasks, mechanisms for flagging factual uncertainty, and routines for the model's own self-assessment. Such artifacts function as organizational control devices applied not to people but to algorithmic systems embedded in the process of intellectual work, which conceptually links them to the accountability and bureaucratization mechanisms studied by classical organization theory, now displaced to a boundary zone between human and machine (Faraj, Pachidi, & Sayegh, 2018).

The research problem guiding this article stems from a twofold gap. First, the Management literature on artificial intelligence in organizations has concentrated predominantly on the effects of AI use on decision-making processes, productivity, and the configuration of work (Von Krogh, 2018), with less attention to how AI systems themselves report on, interpret, and legitimize—or delegitimize—governance instruments designed to regulate them. Second, there are virtually no studies that systematically and simultaneously compare how different families of foundation models (proprietary and open, Western and non-Western) respond to the same normative instrument, which precludes any generalization about the homogeneity or heterogeneity of so-called "alignment philosophies" when confronted with demands for academic auditability. This gap is relevant because protocols of this kind have increasingly been proposed, spontaneously, by graduate researchers as a practical response to the absence of formal guidelines from educational institutions themselves, and their eventual institutionalization by graduate programs presupposes knowledge about the reliability and consistency of the self-assessments that the systems produce about themselves.

The relevance of the topic for the field of Management lies at the intersection of three established fields: algorithmic governance and the accountability of intelligent systems embedded in organizational processes (Diakopoulos, 2016); the theory of organizational trust applied to nonhuman agents (Mayer, Davis, & Schoorman, 1995; Lee & See, 2004); and organizational learning in contexts of accelerated technological change (Argyris & Schön, 1978). Understanding how generative AI systems themselves discursively construct their adherence to, resistance to, or simulation of compliance with a control protocol is therefore a matter of innovation management

and institutional policy design, not merely a technical problem of prompt engineering. Additionally, the fact that the corpus analyzed is produced by the very systems under analysis introduces a little-explored methodological dimension—self-referentiality—that speaks directly to classical discussions in Management about self-assessment, impression management, and organizational legitimation (Goffman, 1959; DiMaggio & Powell, 1983).

In light of this scenario, the general objective of this article is to analyze, through Content Analysis, the meanings that six different foundation models attribute, in self-produced technical-critical assessments, to the operational feasibility, cost-benefit, and application risks of a user–AI interaction governance protocol aimed at master's and doctoral students in Management, identifying convergences, divergences, and patterns of self-representation among these systems.

To that end, the article is structured into six sections in addition to this introduction. The second section presents the theoretical framework, organized into three subsections that discuss, respectively, algorithmic governance and accountability in organizations, trust and heuristics in the relationship between humans and intelligent systems, and institutional isomorphism applied to AI-mediated organizational learning. The third section details the Content Analysis methodological path adopted, including the description of the corpus, the selection criteria, the categorical grid, and the scientific rigor criteria observed. The fourth section presents and discusses the emergent results, confronting them with the Management literature. The fifth section offers the concluding remarks, returning to the general objective, noting the study's theoretical and methodological limitations, and proposing an agenda for future research. Finally, the sixth section gathers the bibliographic references cited throughout the text.

2. THEORETICAL FRAMEWORK

2.1 Algorithmic governance and accountability in generative AI systems

The incorporation of artificial intelligence systems into organizational processes reconfigures the classical distribution of authority and control described by Mintzberg (1979) in characterizing the mechanisms of work coordination. When the "worker" whose activity is to be coordinated is a nondeterministic statistical model, the standardization of processes—one of Mintzberg's five coordinating mechanisms—takes on an unprecedented form: what is standardized is not human behavior but the interaction between human and algorithm, by means

of instructions embedded in the prompt itself. Diakopoulos (2016) describes this type of requirement as algorithmic accountability, that is, the obligation of automated systems to account for their decision criteria in a manner auditable by third parties. The User–GenAI Pact analyzed in this study can be read, in these terms, as an artisanal accountability device, created by the user-researcher in the absence of institutional regulation, which brings it close to the notion of "adaptive governance"—emerging regulatory arrangements that precede and inform formal regulation (Faraj, Pachidi, & Sayegh, 2018). However, Faraj et al. (2018) caution that learning algorithms operate as "opaque actors" whose internal logic is not fully accessible even to the system itself, which imposes a structural limit on any protocol that demands from the model a reliable self-audit of its own internal processes—a central tension that this article explores empirically. Along these lines, Simon (1957) had already warned, in formulating the concept of bounded rationality, that every decision maker—human or artificial—operates under constraints of information and processing capacity that make full rationality unattainable; protocols such as the Pact can be interpreted as heuristics of organizational satisficing, that is, attempts to make manageable, through simplifying rules, a working relationship intrinsically marked by radical uncertainty about the behavior of the regulated system.

2.2 Trust, heuristics, and the paradox of self-assessment in intelligent systems

The organizational trust literature offers a complementary lens for understanding why researchers turn to rigid protocols for interacting with AI. Mayer, Davis, and Schoorman (1995) define trust as one party's willingness to be vulnerable to the actions of another, based on the expectation of that counterpart's ability, benevolence, and integrity; when the counterpart is a nonhuman system, Lee and See (2004) propose the concept of trust calibration, according to which performance is optimized when the user's degree of trust corresponds to the system's actual capability—neither undertrust, which underuses the tool, nor overtrust, which exposes the user to undetected errors. The design of the Pact, by requiring explicit flags of factual uncertainty and periodic coherence "decay" tests, constitutes a procedural attempt to force this calibration. Nevertheless, the literature on human–algorithm interaction describes heterogeneous and at times contradictory reactions to fallible systems: Dietvorst, Simmons, and Massey (2015) identify so-called algorithm aversion, whereby users abandon algorithmic tools as soon as they see them err, even when the tools outperform the average human, whereas Logg, Minson, and Moore (2019)

document the opposite phenomenon, algorithm appreciation, in which automated judgments are preferred to equivalent human judgments. Added to this is the epistemological problem identified by Goffman (1959) in his analysis of the "presentation of self": every account an actor produces of his or her own performance is simultaneously performance and information, so that a language model's self-assessment of its own conformity to a protocol—as demanded by the Pact—cannot be taken as independent evidence of its actual conduct but as a discursive genre of impression management, which theoretically grounds the hypothesis of self-referential bias explored in the results section.

2.3 Institutional isomorphism and organizational learning in AI-mediated scientific knowledge management

Finally, the adoption—real or intended—of protocols such as the User–GenAI Pact by graduate programs speaks to DiMaggio and Powell's (1983) theory of institutional isomorphism, according to which organizations embedded in the same field tend to converge toward similar structures and practices through coercive, mimetic, and normative mechanisms, regardless of their proven effectiveness. The rapid dissemination of "pacts" and "meta-prompts" among researchers can be interpreted as mimetic isomorphism at an early stage, motivated by widespread uncertainty about best practices for the academic use of GenAI. Argyris and Schön (1978) contribute the distinction between single-loop learning, which corrects deviations without questioning the premises underlying action, and double-loop learning, which reexamines those premises; protocols such as the Pact operate predominantly in the former register, standardizing responses to the model's recurrent errors (hallucination, complacency, loss of context) without altering the cognitive architecture of the regulated system. Complementarily, Teece (2007) situates the capacity for sensing, seizing, and reconfiguring at the core of organizational dynamic capabilities in the face of technological change; from this perspective, the autonomous development of governance protocols by individual researchers constitutes an emergent and not yet institutionalized form of dynamic capability at the micro level, whose effectiveness depends, as this study will seek to demonstrate, on the consistency with which the regulated systems themselves recognize and execute it. Pettigrew (1987) reinforces that all organizational change must be understood in the context–process–content triad: the content of the Pact (its clauses) cannot be evaluated in isolation from the process by which researchers adopt it turn by turn, nor

from the institutional context of normative absence in which it emerges—a theoretical articulation that guides the processual reading of the findings presented below.

3. METHODOLOGY

A qualitative approach of an interpretive nature was adopted, operationalized through Content Analysis in the tradition of Bardin (2016), complemented by Krippendorff's (2018) reliability criteria and Bauer's (2002) procedural systematization. The methodological choice is justified by the object itself: the material is discursive, unstructured text produced in natural language, for which Content Analysis offers the most appropriate path from an impressionistic reading to systematized and replicable inferences about the conditions under which the texts were produced (Bardin, 2016).

3.1 Analytical corpus and selection criteria

The corpus consists of twelve texts, organized into six documentary units (one per model), each containing two technical-critical assessments produced in response to two sequential, complementary prompts: the first requested an evaluation of the operational feasibility and cost-benefit of a governance protocol—the "Pact (Meta-Prompt) for User–GenAI Interaction, version T13"—for master's and doctoral students in Management; the second requested a predictive comparative analysis of how different families of foundation models would tend to behave in response to the same protocol. The six models selected—Google Gemini, Alibaba Qwen, DeepSeek, Meta Llama, OpenAI ChatGPT, and Anthropic Claude—were chosen by purposive sampling, according to the criterion of representativeness of the main commercially available families of foundation models, deliberately including variation between proprietary and open architectures and between distinct geographic origins (United States and China), so as to maximize the discursive heterogeneity of the corpus. The material totals approximately 26,500 words in Portuguese, collected in August 2026.

3.2 Units of analysis and categorical grid

The recording unit was defined as the paragraph or structured item (clause, stage of the research cycle, or commented model family), and the context unit as the complete assessment by each model for each of the two prompts, preserving the argumentative coherence required by

Bardin (2016). The categorical grid combined two complementary procedures. A priori categories were derived from the structure of the object itself—the Pact's native tripartition into capabilities that are "directly executable," "functionally simulated," and "not guaranteeable," and the segmentation by stages of the scientific research cycle (problem formulation, literature review, theoretical framework, method, data collection, analysis, discussion, conclusion, writing). A posteriori categories emerged from the process of floating reading and successive coding, making it possible to capture themes not anticipated by the Pact's own structure, such as the patterns of self-referencing among models.

3.3 Coding procedure

The three chronological poles proposed by Bardin (2016) were followed: (i) pre-analysis, with a full floating reading of the corpus, formulation of initial hypotheses, and constitution of the definitive corpus; (ii) exploration of the material, with systematic coding of the recording units in a categorical spreadsheet cross-referencing source model, prompt (1 or 2), and thematic category; and (iii) treatment of results, inference, and interpretation, in which thematic frequencies and recurrences were confronted with the theoretical framework. Coding was performed iteratively, with at least two rounds of full rereading of the corpus to verify the stability of the categories and adjust ambiguous units—a procedure that approximates what Krippendorff (2018) calls stability verification in individual coding, in the absence of a second human coder.

3.4 Scientific rigor criteria

The qualities of good categorization proposed by Bardin (2016) were observed: exhaustiveness (coverage of all pertinent material), representativeness (a sample that reflects the universe of discourse on the topic), homogeneity (categories built according to a single classificatory principle), pertinence (fit of the categories to the object and objectives of the research), and mutual exclusivity (no overlap between categories). Additionally, the semantic validity check suggested by Krippendorff (2018) was employed, consisting of verifying that the category labels in fact corresponded to the coded excerpts, through systematic return to the source text before any inference. It is acknowledged, as a methodological limitation assumed here and revisited in the final section, that formal intercoder reliability indices (such as Krippendorff's alpha) could not be calculated, given that the coding was conducted by a single analyst.

4. RESULTS AND DISCUSSION

The coding of the corpus produced three central analytical categories, presented below in order of recurrence: (i) contested stratification of declared technical capabilities; (ii) asymmetry of costs and benefits across the scientific research cycle; and (iii) self-referential bias in the comparison among model families.

4.1 Contested stratification of declared technical capabilities

All six models reproduced, with lexical variations, the Pact's own tripartition among directly executable behaviors (formatting, turn markers, checklists), merely functionally simulated behaviors (coherence "decay" tests, self-assessment of ethical adherence), and capabilities that cannot be guaranteed by architectural design (persistent memory across sessions, verifiable introspection into their own generation processes). This structural consensus is relevant because it indicates that the limitation of internal auditability is not an isolated admission but a convergent trait across competing architectures—a finding that empirically corroborates the theoretical warning of Faraj, Pachidi, and Sayegh (2018) about the constitutive opacity of learning systems, even when reported by the systems themselves. One of the assessments explicitly named this limitation a "structural self-reference problem," stating that asking the evaluated system to apply to itself the instrument that evaluates it turns every conformity classification into reported behavior rather than external verification—a formulation that converges directly with the Goffmanian reading of self-assessment as discursive performance (Goffman, 1959), adopted in this study's theoretical framework. However, divergences arose regarding the degree of reliability attributed to the intermediate layer of "functional simulation": models with greater emphasis on explicit reasoning chains reported a lower risk of "epistemic theater," whereas others described the same mechanism as producing an "illusion of control" that the user, under deadline pressure, may treat as more reliable than it actually is.

4.2 Asymmetry of costs and benefits across the scientific research cycle

The second category emerged from the cross-coding of model and stage of the research cycle. There was substantial convergence around an inverted-U pattern: the perceived net gain is high in the stages of formulating objectives and hypotheses, methodological design, and final textual revision, and declines in the stages of primary data collection and analysis, in which the

assessments repeatedly pointed to high operational cost, risk of exposing sensitive data, and the need for specialized tools external to the language model itself. This pattern is interpretable in light of Argyris and Schön's (1978) distinction between single-loop and double-loop learning: the protocols analyzed prove effective at correcting superficial, recurrent deviations (ambiguous prompts, tautological hypotheses) but do not alter the model's structural limitation in tasks that require direct, verifiable contact with the empirical reality of the research field. An additional finding, not anticipated in the a priori categories, was the so-called "scaffolding effect followed by dependency paradox," described by one of the models: use of the protocol would increase the researcher's prompt-formulation competence in the first months of use, but, once this threshold is surpassed, the growing delegation of cognitive tasks to the system would tend to reduce the deliberate practice of autonomous critical judgment—a finding that speaks to the dynamic capabilities literature (Teece, 2007) by suggesting that individual algorithmic governance, in order to generate sustainable organizational learning, would need to be designed as a temporary rather than permanent device.

4.3 Self-referential bias in the comparison among model families

The most robust category emerged from the corpus's second prompt, in which each model was invited to predict the behavior of the other families in response to the same protocol. In five of the six documentary units, the responding model classified itself, or the architectural family closest to its own, in the most favorable category of formal adherence, critical rigor, and resistance to complacency, while attributing to competing models a greater propensity toward "ritual compliance," "complacency," or "scope degradation" over the course of the interaction. This systematic pattern of discursive self-favoritism constitutes robust empirical evidence in support of the hypothesis, derived from the Goffmanian framework and the organizational trust literature, that the self-reports of algorithmic systems operate as impression management rather than as neutral measurement of performance (Goffman, 1959; Mayer, Davis, & Schoorman, 1995). The finding has direct implications for institutional governance: if each system tends to preferentially validate its own conformity, any decision to adopt protocols such as the Pact based exclusively on the self-assessment of the model in use incurs systemic confirmation bias, reinforcing the recommendation, also cited in the assessments analyzed, of external human verification as a condition of institutionalization—which brings the finding close to Lee and See's (2004) notion

of trust calibration, according to which appropriate trust in automated systems depends on sources of evaluation independent of the evaluated system itself.

5. CONCLUDING REMARKS

This article aimed to analyze the meanings that six different foundation models attribute, in self-produced technical-critical assessments, to the operational feasibility, cost-benefit, and risks of a user–AI interaction governance protocol aimed at master's and doctoral students in Management. This objective was achieved through Content Analysis of a corpus of twelve texts, which made it possible to identify three central analytical categories. The first revealed a structural consensus among the models regarding the impossibility of a fully reliable self-audit, albeit with divergences in the degree of reliability attributed to the intermediate mechanisms of "functional simulation" of compliance. The second demonstrated an asymmetric cost-benefit pattern across the research cycle, favorable to the stages of conceptual and methodological structuring and unfavorable to the stages of direct contact with empirical data, complemented by the unexpected finding of a possible cognitive dependency effect resulting from prolonged use of the protocol. The third, and most robust, evidenced a systematic pattern of self-referential bias, in which the models tended to evaluate themselves, or architecturally similar systems, more favorably than their competitors, corroborating the theoretical reading of algorithmic self-assessment as an exercise in impression management rather than neutral measurement of performance.

The study has theoretical and methodological limitations that should be acknowledged. Methodologically, the coding was conducted by a single analyst, without human intercoder verification, which precludes the calculation of formal statistical reliability indices, although stability procedures through successive rereading were adopted as a partial safeguard. The corpus is exclusively declarative in nature—consisting of textual accounts that the models produce about themselves—rather than actual behavioral observation of the protocol's execution in real, prolonged interactions, so that the findings describe discursive representations, not verified performance. Theoretically, the analysis focused on a single governance artifact and a single disciplinary domain of application (Management research), which limits the generalization of the identified patterns to other protocols or fields of knowledge. Additionally, the corpus is temporally

delimited to August 2026, a period of rapid technical evolution of foundation models, so that the reported behaviors may not hold in future versions of the same systems.

As an original contribution, the study introduces into the Management literature systematized evidence of the phenomenon of self-referential bias in the self-reports of generative artificial intelligence systems, shifting the discussion on organizational trust in intelligent systems (Mayer, Davis, & Schoorman, 1995; Lee & See, 2004) to a still little-explored terrain: the simultaneous comparison of multiple competing architectures in response to the same normative instrument. It also contributes practically by providing graduate programs with evidence that the institutionalization of protocols for the use of generative AI should not rest exclusively on the self-assessments produced by the systems themselves but requires independent mechanisms of human verification, in line with Lee and See's (2004) recommendation of trust calibration.

As an agenda for future research, we suggest: (i) controlled experimental studies, with real and simultaneous application of the same protocol to the same model families, making it possible to confront the discursive predictions identified here with actual behavioral performance; (ii) extension of Content Analysis to corpora from other disciplinary areas, testing the generalizability of the self-referential bias pattern; (iii) incorporation of human intercoder coding, with calculation of formal reliability indices, in future replications of this design; and (iv) longitudinal investigation of the "dependency paradox" identified in the cost-benefit category, through panel studies with graduate researchers who are real users of protocols for governing interaction with generative AI.

REFERENCES

Argyris, C., & Schön, D. A. (1978). Organizational learning: A theory of action perspective. Addison-Wesley.

Bardin, L. (2016). Análise de conteúdo [Content analysis] (L. A. Reto & A. Pinheiro, Trans.). Edições 70.

Bauer, M. W. (2002). Análise de conteúdo clássica: Uma revisão [Classical content analysis: A review]. In M. W. Bauer & G. Gaskell (Eds.), Pesquisa qualitativa com texto, imagem e som: Um manual prático [Qualitative researching with text, image and sound: A practical handbook] (pp. 189–217). Vozes.

Diakopoulos, N. (2016). Accountability in algorithmic decision making. Communications of the ACM, 59(2), 56–62. https://doi.org/10.1145/2844110

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General, 144(1), 114–126. https://doi.org/10.1037/xge0000033

DiMaggio, P. J., & Powell, W. W. (1983). The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. American Sociological Review, 48(2), 147–160. https://doi.org/10.2307/2095101

Faraj, S., Pachidi, S., & Sayegh, K. (2018). Working and organizing in the age of the learning algorithm. Information and Organization, 28(1), 62–70. https://doi.org/10.1016/j.infoandorg.2018.02.005

Goffman, E. (1959). The presentation of self in everyday life. Doubleday.

Krippendorff, K. (2018). Content analysis: An introduction to its methodology (4th ed.). Sage.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 46(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. Organizational Behavior and Human Decision Processes, 151, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. Academy of Management Review, 20(3), 709–734. https://doi.org/10.2307/258792

Mintzberg, H. (1979). The structuring of organizations: A synthesis of the research. Prentice-Hall.

Pettigrew, A. M. (1987). Context and action in the transformation of the firm. Journal of Management Studies, 24(6), 649–670. https://doi.org/10.1111/j.1467-6486.1987.tb00467.x

Simon, H. A. (1957). Administrative behavior: A study of decision-making processes in administrative organization (2nd ed.). Macmillan.

Teece, D. J. (2007). Explicating dynamic capabilities: The nature and microfoundations of (sustainable) enterprise performance. Strategic Management Journal, 28(13), 1319–1350. https://doi.org/10.1002/smj.640

Von Krogh, G. (2018). Artificial intelligence in organizations: New opportunities for phenomenon-based theorizing. Academy of Management Discoveries, 4(4), 404–409. https://doi.org/10.5465/amd.2018.0084