Journal of Multidisciplinary Engineering Science and Technology (JMEST)
ISSN: 2458-9403
Vol. 12 Issue 10, October - 2025

Managing Shadow AI Adoption in Corporate Environments: A Risk-Based Governance Model and Controlled Classification Experiment

Ozuomba Simeon

TETFUND Center of Excellence in Computational Intelligence Research,
University of Uyo
Akwa Ibom State , Nigeria
simeonozuomba@uniuyo.edu.ng, simeonoz@yahoo.com

Abstract

Generative artificial intelligence has entered ordinary corporate work through public chatbots, embedded copilots, browser extensions, automation agents and AI-enabled software services. This diffusion has created a governance problem: employees and business units can now use powerful AI capabilities without formal approval, security assessment, procurement review or compliance oversight. This paper develops the Shadow AI Risk-Based Governance Model (SAI-RGM) and operationalizes it through a controlled 960-record classification experiment. The model combines Shadow IT theory, enterprise risk management, responsible AI governance, information-security controls and AI risk-management standards. It has six components: discovery and visibility, risk classification, acceptable-use zoning, control selection, continuous monitoring and adaptive governance feedback. The experiment encodes five governance attributes–data sensitivity, output impact, tool-trust risk, system-integration level and oversight weakness–and compares five classifiers for predicting governance zones. Within the controlled taxonomy-based dataset, the best-performing support vector machine achieved 98.96% hold-out accuracy, 99.23% balanced accuracy and 0.991 macro F1-score, while five-fold cross-validation produced a mean macro F1-score of 0.978. The results demonstrate internal consistency and automation feasibility within the controlled dataset, but they do not constitute independent real-world validation. The study contributes by moving Shadow AI governance from broad policy statements to a practical risk-zoning architecture that supports controlled innovation, stronger accountability and proportionate oversight.

Keywords: Shadow AI; AI governance; generative AI; enterprise risk management; information security; controlled classification experiment; risk-based governance; corporate compliance.

1. Introduction

Artificial intelligence has moved into the ordinary routines of corporate work. Employees use generative AI tools to draft reports, summarize meetings, translate documents, write code, classify complaints, prepare presentations and search large bodies of text. Many of these uses are useful, but they also occur faster than formal governance can respond. Shadow AI describes this unmanaged layer of AI adoption: the use of AI tools, models, plugins, copilots or automation agents by employees, teams or business units without formal authorization, security assessment, procurement approval or compliance review. The phenomenon builds on the older problem of Shadow IT, where employees adopt technologies outside formal IT control when official systems fail to meet practical work demands [1], [2].

The governance difficulty is sharper in AI than in ordinary software use. Generative AI systems do not only store and process data; they transform prompts, infer patterns, generate persuasive outputs and may connect to documents, code repositories, cloud services or enterprise applications. AI risk-management frameworks therefore emphasize governance, mapping, measurement and management of AI risk across the lifecycle [3]. The management-system perspective in ISO/IEC 42001 and the AI risk guidance in ISO/IEC 23894 also treat AI as an organizational control problem rather than a purely technical tool [4], [5].

Shadow AI introduces immediate risks for confidentiality, intellectual property, cybersecurity, regulatory compliance and decision accountability. Large-language-model security guidance identifies prompt injection, sensitive-information disclosure and insecure output handling as major risks in AI-enabled applications [6]. The European Union's AI Act also reflects a risk-based regulatory logic in which obligations increase when AI systems affect sensitive domains, legal rights

or safety-critical processes [7]. These developments make blanket prohibition an inadequate response. A ban may reduce visible AI use while increasing hidden use through personal accounts, private devices or unmonitored browser tools.

The research gap is clear. Shadow IT literature explains why employees adopt unauthorized technologies, but it does not fully capture prompt-based data exposure, probabilistic outputs, hallucination, model opacity, insecure plugins and AI-assisted decision influence. Responsible AI governance literature provides principles and structures, but it usually assumes that AI systems are formally developed, procured or deployed. Corporate organizations need a model that can identify Shadow AI use, classify its risk, apply proportionate controls and still preserve legitimate productivity gains.

The study addresses this gap by developing SAI-RGM and testing its operational logic through a controlled classification experiment. The research question is: how can corporate organizations manage Shadow AI adoption through a risk-based governance model that balances innovation, security, compliance and accountability? The objectives are to define Shadow AI in corporate environments, identify its main risk categories, develop a risk-based governance model, operationalize the model through governance-zone prediction and specify implementation roles, controls and monitoring indicators.

2. Literature Review and Conceptual Background

2.1 Shadow IT and the Emergence of Shadow AI

Shadow IT refers to information systems, applications and digital services used within an organization without formal authorization from the IT function [1]. It often emerges because employees need speed, flexibility and functionality that sanctioned systems do not provide. Later Shadow IT research shows that unauthorized tools may become tolerated or formally governed when they prove useful and when the organization creates appropriate control arrangements [2].

Shadow AI follows the same user-driven logic, but it changes the risk surface. In Shadow IT, the employee usually adopts a tool. In Shadow AI, the employee also becomes a prompt designer, data provider, output reviewer and decision intermediary. The same AI tool may be harmless when used to improve a public paragraph and unacceptable when used to summarize employee medical records or draft legal advice. Governance must therefore focus on use cases, data, outputs and human oversight rather than tool names alone.

Table 1. Comparison between Shadow IT and Shadow AI

Dimension	Shadow IT	Shadow AI
Core concern	Unauthorized software, hardware, cloud services or user-managed applications.	Unauthorized AI tools, models, copilots, browser extensions, agents, APIs and AI-enabled services.
Typical exposure	Licensing, integration, security, duplication and compliance risk.	Data leakage, prompt exposure, hallucination, bias, insecure outputs, model opacity and decision accountability.
System behavior	Mostly deterministic software functions.	Probabilistic, context-sensitive and sometimes non-repeatable outputs.
Employee role	User and adopter of a tool.	Prompt designer, data provider, output reviewer and decision intermediary.
Governance focus	Asset visibility, access control and procurement compliance.	Use-case visibility, data protection, output reliability, oversight and lifecycle monitoring.
Control logic	IT asset management and access restrictions.	Risk zoning, data controls, human review, vendor assurance and adaptive governance.

2.2 Drivers of Shadow AI Adoption

Shadow AI adoption is usually driven by a mixture of business pressure and tool convenience. Employees turn to unofficial AI systems when approved tools are unavailable, slow, difficult to use or perceived as weaker than public alternatives. Other drivers include unclear policy, weak communication from IT and compliance teams, peer influence, curiosity, fear of being judged for needing support and pressure to deliver faster work.

These drivers are important because they show that Shadow AI is not always simple misconduct. It can reveal unmet organizational capability needs. A department that repeatedly uses public AI tools to summarize long documents may need an approved enterprise summarization tool. A coding team that relies on personal AI copilots may need an approved development assistant with secure repository controls. Governance is stronger when it converts useful hidden practices into visible and controlled practices.

2.3 Risk Categories Associated with Shadow AI

The first risk category is data confidentiality. Employees may paste confidential documents, source code, customer records, personnel information, legal material, financial data or trade secrets into unauthorized AI tools. A second category is cybersecurity. Prompt injection, insecure plugins, malicious outputs and unsafe code generation can create vulnerabilities in AI-enabled workflows [8].

A third category is decision quality. Generative AI outputs can be fluent but wrong, incomplete or fabricated. Explainable AI research shows that transparency and interpretability are important when automated outputs influence consequential decisions [9]. Ethical AI literature also stresses beneficence, non-maleficence, autonomy, justice and explicability as basic principles for responsible AI use [10].

A fourth category is accountability. When employees act on AI-generated recommendations without documenting the tool, data, prompt, review process or human decision owner, responsibility becomes unclear. Responsible AI governance research emphasizes the need for structural, relational and procedural mechanisms that assign accountability across organizations [11]. Case-based research on AI governance also shows that firms need explicit practices for overcoming barriers and producing trustworthy AI outcomes [12]. Internal algorithmic auditing and human-centered AI principles further support oversight, review and accountability [13], [14].

A fifth category is societal and language-model harm. Large language models can create risks involving biased outputs, information hazards, misinformation, malicious use and unsafe human-computer interaction [15], [16]. These risks do not disappear when AI is used informally inside a company; they become harder to detect. Foundation-model research also shows that scale and generality expand both usefulness and risk [17].

2.4 Existing AI Governance Approaches

Public AI principles from the OECD, UNESCO and the European Commission High-Level Expert Group on AI have encouraged human-centric, trustworthy and rights-respecting AI [18], [19], [20]. These frameworks are useful, but employees need operational rules that tell them what they may do with particular tools, data types and outputs. Enterprise governance also depends on information-security controls. ISO/IEC 27001 and ISO/IEC 27002 provide management-system and control guidance that can be connected to AI use, especially for access control, supplier relationships, monitoring, incident response and data protection [21], [22].

The SAI-RGM model builds on these foundations by translating high-level principles into a use-case governance cycle. It does not ask organizations to choose between innovation and control. It gives them a way to sort AI use by risk and apply the right level of governance response.

3. Theoretical Foundation

The model integrates five theoretical and practice foundations: Shadow IT theory, enterprise risk management, responsible AI governance, information-security governance and AI management-system standards. The combination is necessary because Shadow AI is a socio-technical issue. It involves employee behavior, tool access, data movement, third-party platforms, model reliability, policy compliance and business incentives.

Table 2. Theoretical foundations and their contribution to SAI-RGM

Foundation	Relevance to Shadow AI	Contribution to the model
Shadow IT theory	Explains hidden technology adoption and the movement from unauthorized to governed use.	Supports discovery, visibility and employee disclosure mechanisms.
Enterprise risk management	Provides a structured way to compare likelihood, impact, risk appetite and treatment priorities.	Supports risk scoring, zone assignment and proportional control selection.
Responsible AI governance	Emphasizes accountability, transparency, human oversight, fairness, safety and lifecycle responsibility.	Supports acceptable-use zoning, review mechanisms and role accountability.
Information-security governance	Addresses confidentiality, integrity, access control, monitoring, supplier risk and incident response.	Supports technical controls, data controls and continuous monitoring.
AI management-system standards	Promote formal governance, documented procedures, risk management and continual improvement.	Supports adaptive governance feedback and periodic policy review.

Employee behavior research also matters because Shadow AI is shaped by workarounds, compliance motivation and human-AI interaction. Workaround studies show that employees may depart from formal procedures when official processes do not fit practical work demands [23], [24]. Information-security policy research shows that compliance depends on awareness, perceived benefits, costs, motivation, perceived mandatoriness and organizational climate rather than punishment alone [25], [26], [27], [28]. Human-AI interaction research further shows that AI systems require clear feedback, oversight, uncertainty communication and support for human control [29]. Technology acceptance and user-response theories explain why employees adopt, resist or adapt digital tools when they face pressure, uncertainty or perceived usefulness [30], [31].

4. Methodology

4.1 Research Design

The study uses a design that combines conceptual model development with a controlled classification experiment. Conceptual model development is appropriate because Shadow AI is an emerging organizational phenomenon whose governance is still fragmented across information security, AI ethics, legal compliance and business-risk practice. The controlled experiment operationalizes the model by translating governance attributes into structured data and testing whether classifiers can reproduce zone assignments consistently.

The experiment functions as an internal consistency and operational feasibility test. It demonstrates that the governance attributes can be encoded, scored and learned by standard classifiers, while external generalization requires independently labeled real-world Shadow AI use cases from multiple organizations. Because the governance-zone labels were generated from the same risk attributes used for model training, the classification results are interpreted as evidence of structured operationalization and automation feasibility rather than independent field validation.

4.2 Dataset Design and Risk Attributes

A controlled dataset of 960 corporate Shadow AI use-case records was generated from an enterprise risk taxonomy. The taxonomy represents common corporate functions and AI-use patterns, including drafting, summarization, classification, code generation, external communication, analytics, decision support and workflow automation. No personal data, confidential company records or live monitoring data were used.

Each record was described by five ordinal attributes scored from 1 to 4: data sensitivity, output impact, tool-trust risk, system-integration level and oversight weakness. These attributes were selected because they directly affect the harm that can arise from unauthorized AI use.

The records were generated by combining common corporate functions with typical AI-use patterns and governance-relevant risk conditions. The functions included administration, finance, legal, human resources, procurement, customer service, software development, cybersecurity, compliance, operations and executive support. The AI-use patterns included drafting, summarization, translation, classification, code generation, decision support, analytics, external communication and workflow automation. Each generated case was assigned five ordinal risk attributes using the scoring rules in Table 3. Implausible combinations were removed where the stated use case could not reasonably occur in a corporate environment. The final dataset was intentionally imbalanced to reflect triage conditions in which most reported cases fall into controlled or restricted categories rather than equal class groups.

Table 3. Risk attributes used in the controlled classification experiment

Attribute	Score range	Interpretation
Data sensitivity	1-4	Measures whether the input is public, internal, confidential, personal, regulated or strategically sensitive.
Output impact	1-4	Measures whether the AI output is informal, operational, customer-facing, financial, legal, employment-related or safety-relevant.
Tool-trust risk	1-4	Measures approval status, vendor transparency, data-retention terms, security review and contractual assurance.
System-integration level	1-4	Measures whether use is standalone, connected to files, connected to workflows or integrated with enterprise systems/APIs.
Oversight weakness	1-4	Measures the extent of human review, managerial accountability, auditability and documentation.

Table 4. Controlled case-generation procedure

Step	Procedure	Purpose
Corporate-function mapping	Cases were distributed across administration, finance, legal, human resources, procurement, customer service, software development, cybersecurity, compliance, operations and executive support.	Reflects typical functional areas in which employees may adopt AI tools.
AI-use pattern mapping	Cases covered drafting, summarization, translation, classification, code generation, decision support, analytics, external communication and workflow automation.	Captures common generative AI activities in office and technical work.
Risk-attribute assignment	Each case was scored on data sensitivity, output impact, tool-trust risk, integration level and oversight weakness.	Converts qualitative use cases into governance-relevant variables.
Plausibility screening	Implausible combinations were removed before final sampling.	Avoids unrealistic or internally inconsistent records.
Zone labeling	Weighted scoring and the high-risk override were applied to assign Green, Yellow, Orange or Red labels.	Produces a reproducible governance-zone label for controlled experimentation.

4.3 Risk Scoring and Zone Assignment

The governance score for use case i was computed as a weighted sum of the five risk attributes. The weight vector used in the experiment was $w = (1.25,1.15,1.10,0.95,1.05)$, giving slightly higher emphasis to data sensitivity, output impact and tool-trust risk. System-integration level and oversight weakness were also included because they affect operational exposure and accountability risk, but they were weighted slightly lower to avoid over-penalizing low-impact standalone use.

$$R_i = \sum_{j=1}^{5} w_j \, x_{ij} \tag{1}$$

In Equation (1), R_i is the overall risk score for use case i, x_{ij} is the value of attribute j for use case i and w_j is the weight attached to attribute j.

$$R_{\min} = \sum_{j=1}^{5} w_j \,, \qquad R_{\max} = 4 \sum_{j=1}^{5} w_j \tag{2}$$

$$S_i = \frac{R_i - R_{\min}}{R_{\max} - R_{\min}} \tag{3}$$

Equation (3) normalizes the weighted score to a 0-1 exposure scale because each attribute has a minimum value of 1 and a maximum value of 4. The high-risk override was formalized as a Boolean variable to prevent sensitive use cases from being understated by a moderate aggregate score.

$$H_i = \begin{cases} 1, & x_{i1} = 4 \wedge (x_{i2} \geq 3 \vee x_{i4} \geq 3 \vee x_{i5} \geq 3) \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

In Equation (4), x_{i1} represents data sensitivity, x_{i2} output impact, x_{i4} system-integration level and x_{i5} oversight weakness. The final zone assignment is given in Equation (5).

$$Z_i = \begin{cases} \text{Green}, & R_i \leq 10.0 \wedge H_i = 0 \\ \text{Yellow}, & 10.0 < R_i \leq 13.2 \wedge H_i = 0 \\ \text{Orange}, & 13.2 < R_i \leq 17.0 \wedge H_i = 0 \\ \text{Red}, & R_i > 17.0 \vee H_i = 1 \end{cases} \tag{5}$$

For multiclass performance reporting, accuracy was computed from the confusion matrix as follows:

$$\text{Accuracy} = \frac{\sum_{k=1}^{K} C_{kk}}{\sum_{i=1}^{K}\sum_{j=1}^{K} C_{ij}} \quad (6)$$

Macro-F1 was computed as the unweighted mean of the class-specific F1-scores:

$$F1_k = \frac{2P_k R_k}{P_k + R_k}, \qquad \text{Macro-F1} = \frac{1}{K}\sum_{k=1}^{K} F\,1_k \quad (7)$$

In Equations (6) and (7), C is the multiclass confusion matrix, K is the number of governance zones, P_k is class precision and R_k is class recall.

Table 5. Zone thresholds and governance interpretation

Zone	Weighted-score rule	Interpretation
Green	$R_i \leq 10.0$ and $H_i = 0$	Permitted low-risk use under basic AI-use rules and no-sensitive-data restrictions.
Yellow	$10.0 < R_i \leq 13.2$ and $H_i = 0$	Controlled use requiring approved tools, clear data-handling rules and human review.
Orange	$13.2 < R_i \leq 17.0$ and $H_i = 0$	Restricted use requiring formal review and additional technical, legal or compliance controls.
Red	$R_i > 17.0$ or $H_i = 1$	Prohibited, blocked or redesigned because the risk exceeds normal tolerance.

A small sensitivity check was conducted to examine whether modest changes in the weighting approach materially changed the governance distribution. The results in Table 6 show that the distribution remained broadly stable under equal, security-heavy and oversight-heavy weighting approaches, which supports the robustness of the risk-zoning logic for controlled triage.

Table 6. Sensitivity check for alternative weighting approaches

Scenario	Main weighting focus	Green	Yellow	Orange	Red
Equal	All attributes equal.	105	278	386	191
Current	Data sensitivity, output impact and tool-trust risk.	84	369	307	200
Security-heavy	Data and integration exposure.	92	362	315	191
Oversight-heavy	Oversight weakness emphasized.	78	378	312	192

4.4 Experimental Reproducibility Protocol

The dataset was split into 70% training data and 30% hold-out test data using stratified sampling and random seed 42. Five classifiers were compared with a majority-class baseline: multinomial logistic regression, decision tree, random forest, gradient boosting and support vector machine with radial-basis-function kernel. Standardization was applied to logistic regression and SVM. Model reporting included hold-out accuracy, balanced accuracy, macro precision, macro recall, macro F1-score and five-fold cross-validation macro F1-score. The machine-learning algorithms are established methods for classification and ensemble learning [32], [33], [34]. The implementation used Python and scikit-learn [35].

Table 7. Reproducibility protocol for the controlled classification experiment

Item	Specification
Dataset size	960 controlled Shadow AI use-case records.
Input features	Five ordinal attributes scored from 1 to 4.
Label source	Weighted risk taxonomy with high-risk override.

Item	Specification
Train-test split	70/30 stratified split.
Random seed	42.
Validation	Hold-out test set and five-fold stratified cross-validation.
Baseline	Majority-class classifier.
Software	Python with scikit-learn implementation.

5. Development of the Shadow AI Risk-Based Governance Model

SAI-RGM is a cyclical governance model. It begins with visibility rather than punishment. Once a use case is visible, the organization can classify it, assign an acceptable-use zone, select controls, monitor compliance and update governance as tools, regulations and business needs change.

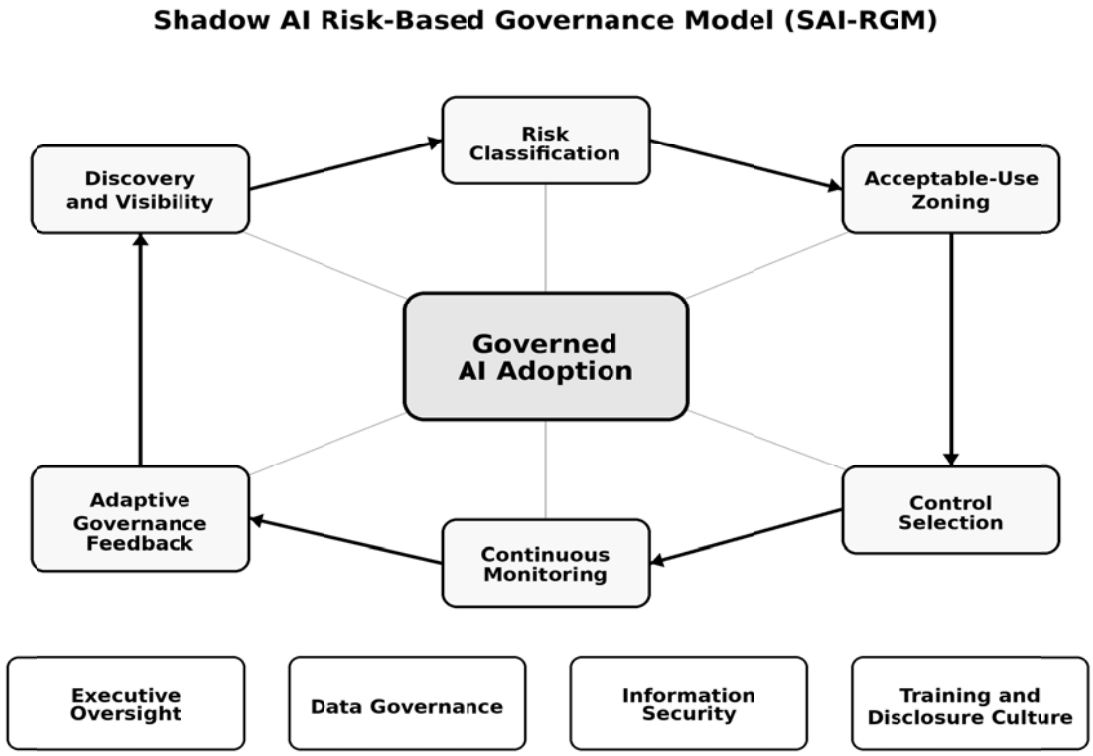

Figure 1. Shadow AI Risk-Based Governance Model (SAI-RGM).

Note: The model operates as a cyclical governance process supported by executive oversight, data governance, information security and training/disclosure culture.

5.1 Discovery and Visibility

Discovery identifies where AI is being used, by whom, for which tasks, with which data and through which tools. It should combine employee disclosure, SaaS discovery, procurement records, browser-extension review, endpoint logs, expense analysis, API traffic and departmental interviews. The output is a living AI use-case inventory.

5.2 Risk Classification

Risk classification assesses the use case rather than the tool alone. The same AI service can be Green for public-language editing and Red for regulated employee or customer decisions. Classification should consider data sensitivity, output consequence, vendor status, integration level and human oversight.

5.3 Acceptable-Use Zoning

Zoning turns risk assessment into practical boundaries. Employees need usable boundaries: what is allowed, what requires approval, what is restricted and what is prohibited.

5.4 Control Selection

Controls should match risk intensity. Low-risk use can be handled with training and clear rules. Higher-risk use requires approved platforms, data-loss prevention, logging, review, legal assessment, vendor assurance and escalation paths.

5.5 Continuous Monitoring

Monitoring should detect unapproved tools, sensitive-data exposure, suspicious prompts, unauthorized integrations, anomalous API activity and repeated policy exceptions. Monitoring should also reveal where employees need safer approved tools.

5.6 Adaptive Governance Feedback

Governance must adapt as models, vendors, laws and work practices change. Feedback from incidents, audits, employee disclosures and business-unit needs should update the inventory, risk zones, control rules and training material.

Table 8. Acceptable-use zones and governance response

Zone	Typical use case	Governance response
Green	Using AI to improve grammar in non-confidential text or brainstorm public information.	Permitted with basic rules, user responsibility and prohibition on sensitive data input.
Yellow	Using an approved enterprise AI tool to summarize internal meeting notes or draft routine internal communication.	Controlled use with approved tools, data-handling rules, human review and logging where appropriate.
Orange	Using AI to analyze customer complaints, contract clauses, code repositories or operational incidents.	Formal review by security, legal, compliance or AI governance leads before use.
Red	Uploading employee medical records, confidential client contracts, regulated decisions or trade secrets into a public AI tool.	Prohibited, blocked or redesigned using approved secure processes.

Table 9. Control matrix by governance zone

Control type	Green	Yellow	Orange	Red
Policy	Basic AI-use rules.	Department AI guidance.	Formal approval.	Prohibition and escalation.
Technical	Access guidance.	Approved platform and logging.	DLP, API control and testing.	Blocking and incident triage.
Data	No sensitive data.	Data-classification rules.	Minimize, mask and retain safely.	No processing unless redesigned.
Human oversight	User review.	Manager review.	Expert review and risk owner.	Use is disallowed.
Vendor assurance	Known service.	Vendor review preferred.	Third-party and contractual review.	Vendor use prohibited.

6. Controlled Classification Results

6.1 Dataset Distribution

The final dataset contained 960 use cases. The distribution was intentionally imbalanced because enterprise triage rarely produces equal numbers of low-, moderate- and high-risk cases. Yellow and Orange cases formed the largest groups, while Green cases were fewer because a corporate governance inventory usually records more than trivial AI use.

Table 10. Governance-zone distribution in the controlled dataset

Zone	Number of use cases	Percentage
Green	84	8.75%
Yellow	369	38.44%
Orange	307	31.98%
Red	200	20.83%

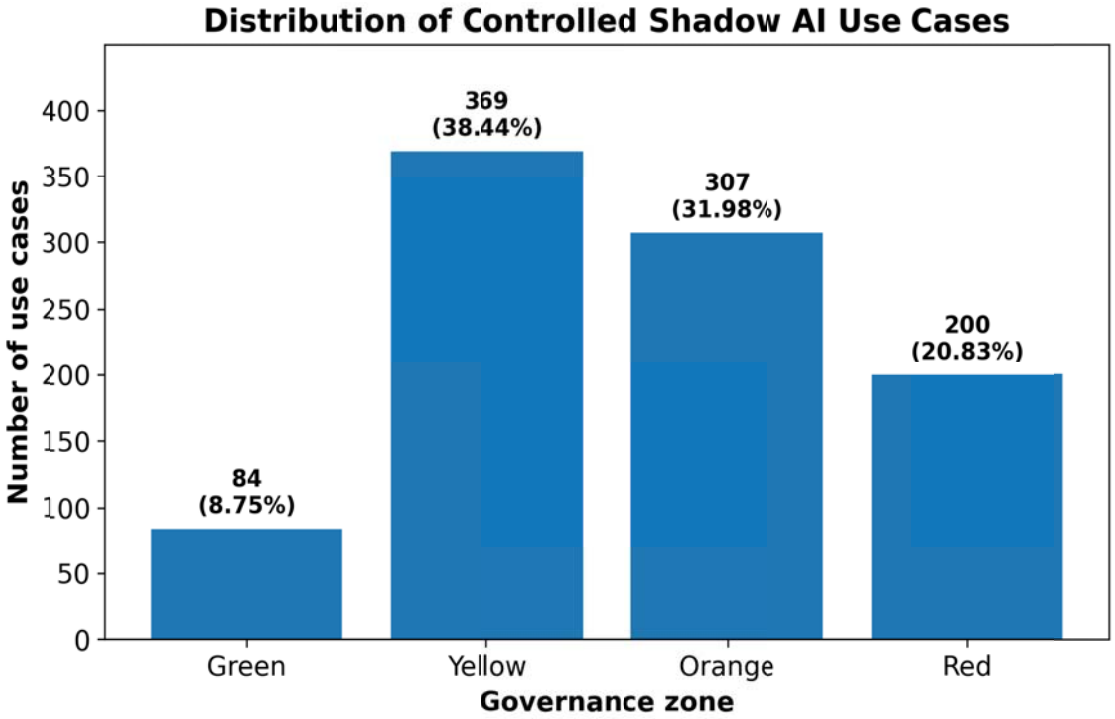

Figure 2. Distribution of controlled Shadow AI use cases by governance zone.

Note: Figure generated from the controlled 960-record Shadow AI use-case dataset used in this study.

6.2 Classifier Performance

Table 11 reports the hold-out and cross-validation results. A deterministic rule-based baseline was included because the controlled labels were generated from the weighted scoring equation and the high-risk override. The rule-based assignment reproduces the taxonomy labels exactly; therefore, the machine-learning comparison is not presented as an attempt to outperform the rule. Rather, it tests whether common classifiers can approximate the governance-zone logic for possible triage automation when embedded in enterprise discovery workflows. The majority-class baseline achieved 38.54% accuracy and 0.139 macro F1-score, confirming that simple majority prediction is not useful. The Support Vector Machine with radial-basis-function kernel achieved the strongest hold-out performance among trainable models, with 98.96% accuracy, 99.23% balanced accuracy and 0.991 macro F1-score. Five-fold cross-validation produced a mean macro F1-score of 0.978, indicating that the result was stable across folds in the controlled dataset. Because the experiment was designed as an operational feasibility test rather than a population-level inference study, the cross-validation mean and standard deviation are reported descriptively; future studies using independently labeled organizational data should apply confidence intervals and paired model-comparison tests.

Table 11. Classifier performance for governance-zone prediction

Model	Accuracy	Balanced accuracy	Macro F1	CV Macro F1 mean	CV SD
Deterministic rule-based scoring	1.0000	1.0000	1.0000	N/A	N/A
SVM (RBF)	0.9896	0.9923	0.9907	0.9781	0.0201
GB	0.9410	0.9120	0.9323	0.8877	0.0494
RF	0.9306	0.9222	0.9256	0.8994	0.0176
MLR	0.9201	0.8822	0.9021	0.9168	0.0119
DT	0.7674	0.7499	0.7616	0.7403	0.0153
Majority	0.3854	0.2500	0.1391	0.1388	0.0005

Note: Deterministic rule-based scoring = weighted risk score with high-risk override; SVM (RBF) = Support Vector Machine with radial-basis-function kernel; GB = Gradient Boosting; RF = Random Forest; MLR = Multinomial Logistic Regression; DT = Decision Tree; Majority = majority-class baseline.

Table 12. Detailed macro metrics for the controlled classification experiment

Model	Macro precision	Macro recall	Macro F1
SVM (RBF)	0.9892	0.9923	0.9907
GB	0.9617	0.9120	0.9323
RF	0.9296	0.9222	0.9256
MLR	0.9319	0.8822	0.9021
DT	0.7759	0.7499	0.7616
Majority	0.0964	0.2500	0.1391

Note: SVM (RBF) = Support Vector Machine with radial-basis-function kernel; GB = Gradient Boosting; RF = Random Forest; MLR = Multinomial Logistic Regression; DT = Decision Tree.

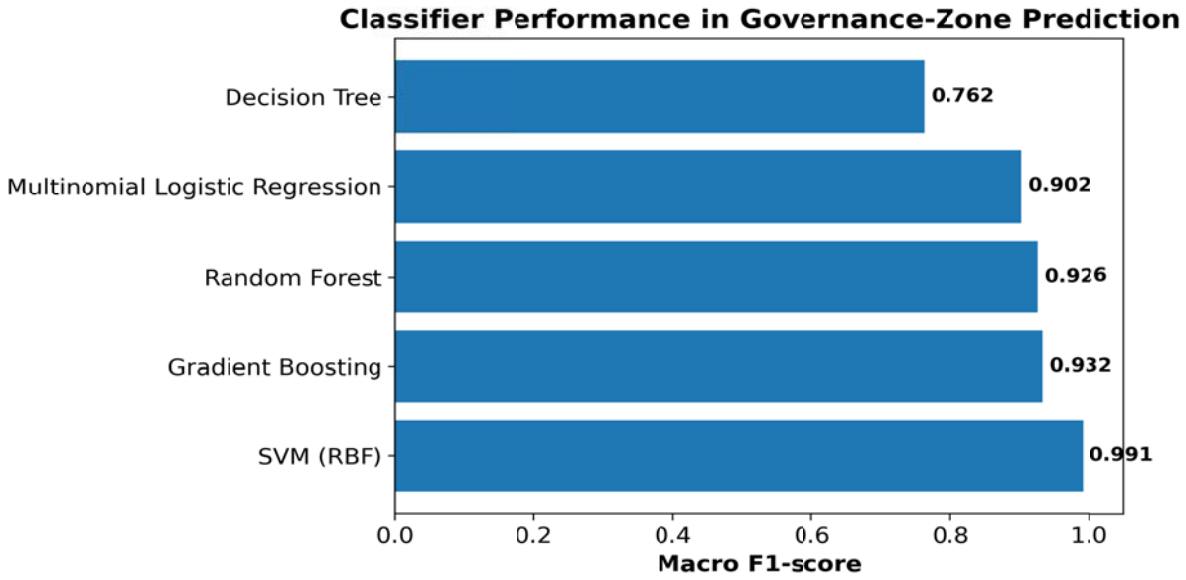

Figure 3. Classifier performance for Shadow AI governance-zone prediction.

Note: Figure generated from the controlled 960-record Shadow AI use-case dataset used in this study.

6.3 Per-Class Results and Error Pattern

The SVM model correctly classified all Green and Red cases in the hold-out set and made only three errors in total. The errors occurred at the Yellow-Orange and Orange-Red boundaries, which is expected because adjacent governance zones differ by degree rather than by completely separate concepts. The classifier should be treated as a triage aid. Final approval of sensitive or consequential AI use remains a governance decision.

Table 13. Per-class performance of the SVM (RBF) classifier

Zone	Precision	Recall	F1-score	Support
Green	1.0000	1.0000	1.0000	25
Yellow	1.0000	0.9910	0.9955	111
Orange	0.9890	0.9783	0.9836	92
Red	0.9677	1.0000	0.9836	60

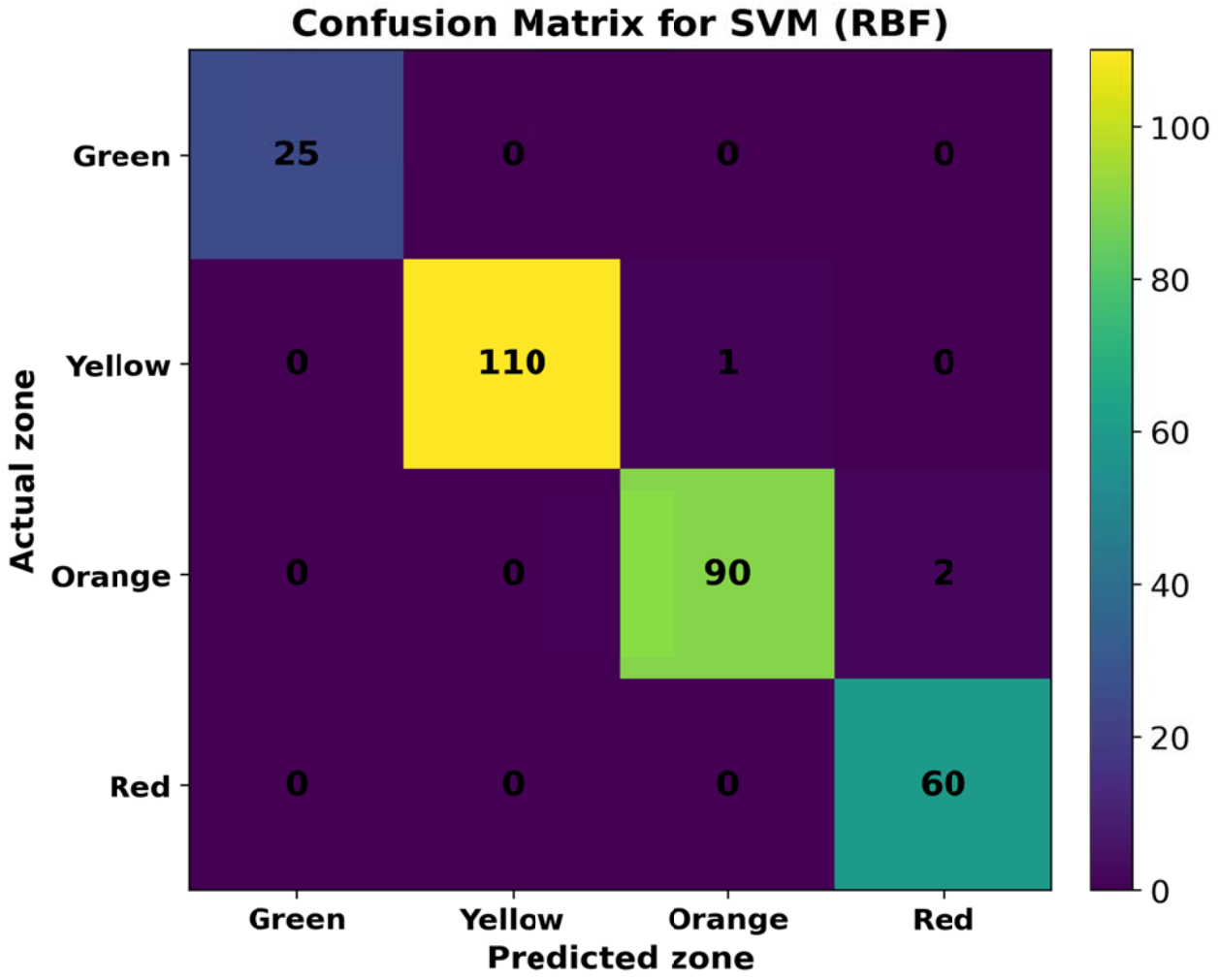

Figure 4. Confusion matrix for the best-performing classifier, SVM (RBF).

Note: Figure generated from the hold-out test set of the controlled Shadow AI use-case dataset.

Table 14. Confusion matrix values for the SVM (RBF) classifier

Actual zone	Predicted Green	Predicted Yellow	Predicted Orange	Predicted Red
Green	25	0	0	0
Yellow	0	110	1	0
Orange	0	0	90	2
Red	0	0	0	60

Table 15. Error analysis for the SVM (RBF) classifier

Error type	No.	Meaning	Governance action
Yellow -> Orange	1	Conservative boundary escalation.	Review before applying stronger controls.
Orange -> Red	2	Conservative escalation of restricted cases.	Human review before blocking or redesign.
Red -> lower zone	0	No high-risk downgrade occurred.	Positive safety signal in the controlled dataset.

7. Implementation Framework

Implementation requires shared ownership. Shadow AI is not owned by one department because its risks cut across technology, law, compliance, procurement, security and business operations. The RACI structure in Table 16 assigns decision rights without removing accountability from business units and employees who use AI in daily work.

Table 16. RACI matrix for Shadow AI governance implementation

Activity	Exec	CIO	CISO	Legal	AIGC	BU	Emp
Define AI risk appetite	A	C	C	C	R	C	I
Maintain AI inventory	I	R	C	C	A	R	I
Classify AI use cases	I	C	C	C	A/R	R	I
Approve high-risk AI use	A	C	C	C	R	C	I
Monitor policy compliance	I	C	R	C	A	C	I
Report unsafe AI use	I	I	C	I	C	C	R

Key: R = Responsible; A = Accountable; C = Consulted; I = Informed; Exec = Executive; AIGC = AI Governance Committee; BU = Business Units; Emp = Employees.

Table 17. Implementation roadmap for SAI-RGM

Phase	Main actions	Expected outputs
1. Governance mandate	Approve policy scope, risk appetite, decision rights and escalation paths.	AI governance charter and executive sponsorship.
2. Shadow AI discovery	Collect employee disclosures, review SaaS use, analyze browser extensions, inspect procurement records and examine relevant traffic patterns.	Initial Shadow AI use-case inventory.
3. Risk classification	Score each use case and assign Green, Yellow, Orange or Red status.	Classified AI inventory and risk heat map.
4. Control deployment	Apply policy, technical, data, human and vendor controls according to zone.	Approved tools, blocked tools, control records and training material.
5. Monitoring and feedback	Track incidents, exceptions, adoption patterns, new tools and user needs.	Updated controls, revised training and improved approved AI services.

Table 18. Practical governance indicators for Shadow AI management

Indicator	Purpose
Number of disclosed AI use cases	Measures visibility and employee reporting.
Percentage of use cases classified by zone	Measures completeness of risk triage.
Number of approved AI tools	Measures safe enablement capacity.
Number of blocked or redesigned Red-zone cases	Measures high-risk exposure reduction.
Sensitive-data alerts involving AI tools	Measures confidentiality-control effectiveness.
Training completion rate	Measures employee readiness.
Exception closure time	Measures governance responsiveness.
Repeat Shadow AI patterns by department	Identifies unmet digital capability needs.

8. Discussion

The findings support a practical interpretation of Shadow AI as a governance paradox. Over-control may suppress useful experimentation and push AI use underground. Under-control may expose the organization to data leakage, unreliable decisions, insecure workflows and regulatory breaches. SAI-RGM addresses this tension by replacing broad prohibition with risk-zoned governance.

The controlled classification experiment shows that the proposed risk attributes can be encoded and used to reproduce governance-zone assignments with high internal consistency within a controlled taxonomy-based dataset. Real-world performance will depend on organizational data quality, sectoral context and expert labeling. The controlled result shows

that the model can be operationalized as a triage mechanism when the organization has defined its risk attributes, thresholds and override rules.

The model also treats Shadow AI as a source of organizational intelligence. Repeated use of unapproved AI tools may reveal where employees need approved summarization, analytics, coding, translation or drafting support. Governance should therefore combine control with service improvement. This view is consistent with responsible-AI engineering work that turns broad principles into system-level governance practices [36], management research that treats AI as both automation and augmentation [37], information-systems scholarship on managing AI as a distinctive organizational technology [38], and studies showing that algorithmic tools reshape workplace control and accountability [39].

9. Theoretical Contributions

The study makes three contributions. First, it extends Shadow IT theory by showing that unauthorized technology adoption in the AI era involves probabilistic output risk, prompt-based data exposure, model opacity and decision accountability. Second, it contributes to responsible AI governance by shifting attention from formally deployed AI systems to hidden, employee-driven AI use. Third, it contributes to enterprise risk management by linking AI use-case attributes to risk zones, controls and governance roles.

The novelty of SAI-RGM lies in its use-case-level governance logic. Rather than treating AI governance only as a tool-approval problem or a general policy problem, the model classifies each AI-enabled activity according to data sensitivity, output consequence, tool trust, integration exposure and oversight weakness. This makes it possible to distinguish low-risk experimentation from sensitive, consequential or poorly supervised AI use. The model therefore extends Shadow IT governance from asset visibility to AI-use-case accountability.

The controlled experiment adds a methodological contribution by showing how governance attributes can be converted into structured data and tested through multiclass classification. This creates a bridge between qualitative governance models and operational risk triage.

10. Practical Implications

For executives, the model provides a way to express AI risk appetite in operational terms. For CIOs and CISOs, it supports approved AI platforms, monitoring, DLP integration and tool review. For legal and compliance functions, it identifies where privacy, contractual, sectoral or regulatory review is required. For business units, it allows legitimate AI use to continue under guardrails. For employees, it replaces vague warnings with concrete examples of permitted, controlled, restricted and prohibited use.

The model is especially useful where organizations cannot immediately approve every AI tool or investigate every use case in depth. It helps them prioritize. Green and Yellow cases can be enabled quickly, while Orange and Red cases receive stronger scrutiny.

11. Limitations and Future Research

The study has limitations. The dataset was controlled and taxonomy-based, making it suitable for testing internal consistency and operational feasibility but not sufficient for claiming universal real-world performance. Because governance-zone labels were generated from the same risk attributes used for classifier training, the experiment should be interpreted as structured operationalization and internal-consistency testing, not as independent empirical validation. Real enterprise data may contain ambiguous use cases, incomplete descriptions, conflicting risk judgments and sector-specific constraints.

Future research should validate SAI-RGM with independently labeled organizational data. A strong next step would be expert labeling by cybersecurity, compliance, legal and business-risk specialists, followed by inter-rater reliability analysis using Cohen's kappa or Fleiss' kappa before model training. Future studies should also test the model in banking, healthcare, energy, education, manufacturing and public administration; compare prohibition-based and risk-zoned governance; and examine employee attitudes toward disclosure and AI monitoring.

12. Conclusion

Shadow AI has become a significant corporate governance issue because AI tools are accessible, useful and often adopted faster than organizations can approve them. Treating the problem only as employee non-compliance misses the

organizational causes of hidden AI use. Employees often turn to unofficial tools because they are trying to solve real work problems.

This study developed SAI-RGM, a risk-based model for managing Shadow AI in corporate environments, and operationalized it through a controlled classification experiment. The model provides a cycle for discovering AI use, classifying risks, assigning acceptable-use zones, selecting controls, monitoring behavior and improving governance over time. The controlled classification component shows that the model's risk attributes can reproduce structured governance-zone assignments within a taxonomy-based dataset. Because governance-zone labels were generated from the same risk attributes used for classifier training, the experiment should be interpreted as structured operationalization and internal-consistency testing rather than independent empirical validation.

The central implication is that organizations should move beyond the question of how to stop Shadow AI. The stronger question is how to make AI use visible, safe, accountable and productive. A risk-zoned governance approach allows low-risk innovation to continue while directing stronger controls toward sensitive, consequential or poorly supervised AI use.

Appendix A. Experimental Parameters

Table 19. Experimental model parameters

Model	Key settings
Majority baseline	Most-frequent class strategy.
Multinomial logistic regression	Standardized features; lbfgs solver; max_iter = 1000; random_state = 42.
Decision tree	Gini criterion; max_depth = 6; random_state = 42.
Random forest	300 estimators; max_depth = 8; balanced class weights; random_state = 42.
Gradient boosting	150 estimators; learning_rate = 0.05; max_depth = 3; random_state = 42.
SVM (RBF)	Standardized features; RBF kernel; C = 10; gamma = scale; random_state = 42.
Validation design	70/30 stratified hold-out split and five-fold stratified cross-validation.

Appendix B. Controlled Dataset Composition

Table 20. Controlled dataset distribution by corporate function

Corporate function	Number of cases	Percentage
Customer service	110	11.46%
Cybersecurity	92	9.58%
Finance	113	11.77%
Human resources	95	9.90%
Legal	82	8.54%
Marketing	95	9.90%
Operations	106	11.04%
Procurement	72	7.50%
Research	100	10.42%
Software engineering	95	9.90%
Total	**960**	**100.00%**

Table 21. Controlled dataset distribution by AI-use pattern

AI-use pattern	Number of cases	Percentage
Analytics	116	12.08%
Automation	118	12.29%
Classification	122	12.71%
Code generation	123	12.81%
Decision support	130	13.54%
Drafting	123	12.81%
External communication	111	11.56%
Summarization	117	12.19%
Total	**960**	**100.00%**

References

[1] Silic, M., & Back, A. (2014). Shadow IT - A view from behind the curtain. Computers & Security, 45, 274-283. https://doi.org/10.1016/j.cose.2014.06.007

[2] Kopper, A., Fürstenau, D., Zimmermann, S., Klotz, S., Rentrop, C., Rothe, H., Strahringer, S., & Westner, M. (2018). Shadow IT and business-managed IT: A conceptual framework and empirical illustration. International Journal of IT/Business Alignment and Governance, 9(2), 53-71. https://doi.org/10.4018/IJITBAG.2018070104

[3] National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. https://doi.org/10.6028/NIST.AI.100-1

[4] International Organization for Standardization. (2023). ISO/IEC 42001:2023 Artificial intelligence - Management system. ISO. https://www.iso.org/standard/42001

[5] International Organization for Standardization. (2023). ISO/IEC 23894:2023 Artificial intelligence - Guidance on risk management. ISO. https://www.iso.org/standard/77304.html

[6] National Institute of Standards and Technology. (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1. https://doi.org/10.6028/NIST.AI.600-1

[7] European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence. Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

[8] Open Worldwide Application Security Project. (2024). OWASP Top 10 for Large Language Model Applications 2023-2024. https://genai.owasp.org/resource/llm-top-10-for-llms-v1-1/

[9] Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82-115. https://doi.org/10.1016/j.inffus.2019.12.012

[10] Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. Harvard Data Science Review, 1(1). https://doi.org/10.1162/99608f92.8cd550d1

[11] Mäntymäki, M., Minkkinen, M., Birkstedt, T., & Viljanen, M. (2022). Defining organizational AI governance. AI and Ethics, 2(4), 603-609. https://doi.org/10.1007/s43681-022-00143-x

[12] Papagiannidis, E., Enholm, I. M., Dremel, C., Mikalef, P., & Krogstie, J. (2023). Toward AI governance: Identifying best practices and potential barriers and outcomes. Information Systems Frontiers, 25(1), 123-141. https://doi.org/10.1007/s10796-022-10251-y

[13] Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 33-44. https://doi.org/10.1145/3351095.3372873

[14] Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe and trustworthy. International Journal of Human-Computer Interaction, 36(6), 495-504. https://doi.org/10.1080/10447318.2020.1741118

[15] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610-623. https://doi.org/10.1145/3442188.3445922

[16] Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P. S., Mellor, J., Glaese, M., Cheng, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., … Gabriel, I. (2022). Taxonomy of risks posed by language models. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 214-229. https://doi.org/10.1145/3531146.3533088

[17] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901. https://papers.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[18] Organisation for Economic Co-operation and Development. (2019). OECD principles on artificial intelligence. https://oecd.ai/en/ai-principles

[19] United Nations Educational, Scientific and Cultural Organization. (2021). Recommendation on the ethics of artificial intelligence. https://unesdoc.unesco.org/ark:/48223/pf0000381137

[20] European Commission High-Level Expert Group on Artificial Intelligence. (2019). Ethics guidelines for trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[21] International Organization for Standardization. (2022). ISO/IEC 27001:2022 Information security, cybersecurity and privacy protection - Information security management systems - Requirements. ISO. https://www.iso.org/standard/27001

[22] International Organization for Standardization. (2022). ISO/IEC 27002:2022 Information security, cybersecurity and privacy protection - Information security controls. ISO. https://www.iso.org/standard/75652.html

[23] Ferneley, E. H., & Sobreperez, P. (2006). Resist, comply or workaround? An examination of different facets of user engagement with information systems. European Journal of Information Systems, 15(4), 345-356. https://doi.org/10.1057/palgrave.ejis.3000629

[24] Alter, S. (2014). Theory of workarounds. Communications of the Association for Information Systems, 34, 1041-1066. https://doi.org/10.17705/1CAIS.03455

[25] Bulgurcu, B., Cavusoglu, H., & Benbasat, I. (2010). Information security policy compliance: An empirical study of rationality-based beliefs and information security awareness. MIS Quarterly, 34(3), 523-548. https://doi.org/10.2307/25750690

[26] Siponen, M., & Vance, A. (2010). Neutralization: New insights into the problem of employee information systems security policy violations. MIS Quarterly, 34(3), 487-502. https://doi.org/10.2307/25750688

[27] Son, J.-Y. (2011). Out of fear or desire? Toward a better understanding of employees' motivation to follow IS security policies. Information & Management, 48(7), 296-302. https://doi.org/10.1016/j.im.2011.07.002

[28] Boss, S. R., Kirsch, L. J., Angermeier, I., Shingler, R. A., & Boss, R. W. (2009). If someone is watching, I'll do what I'm asked: Mandatoriness, control, and information security. European Journal of Information Systems, 18(2), 151-164. https://doi.org/10.1057/ejis.2009.8

[29] Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1-13. https://doi.org/10.1145/3290605.3300233

[30] Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. MIS Quarterly, 27(3), 425-478. https://doi.org/10.2307/30036540

[31] Beaudry, A., & Pinsonneault, A. (2005). Understanding user responses to information technology: A coping model of user adaptation. MIS Quarterly, 29(3), 493-524. https://doi.org/10.2307/25148693

[32] Breiman, L. (2001). Random forests. Machine Learning, 45, 5-32. https://doi.org/10.1023/A:1010933404324

[33] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5), 1189-1232. https://doi.org/10.1214/aos/1013203451

[34] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20, 273-297. https://doi.org/10.1007/BF00994018

[35] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(85), 2825-2830. https://jmlr.org/papers/v12/pedregosa11a.html

[36] Lu, Q., Zhu, L., Xu, X., Whittle, J., Zowghi, D., & Jacquet, A. (2024). Responsible AI Pattern Catalogue: A collection of best practices for AI governance and engineering. ACM Computing Surveys, 56(7), Article 173. https://doi.org/10.1145/3626234

[37] Raisch, S., & Krakowski, S. (2021). Artificial intelligence and management: The automation-augmentation paradox. Academy of Management Review, 46(1), 192-210. https://doi.org/10.5465/amr.2018.0072

[38] Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing artificial intelligence. MIS Quarterly, 45(3), 1433-1450. https://doi.org/10.25300/MISQ/2021/16274

[39] Kellogg, K. C., Valentine, M. A., & Christin, A. (2020). Algorithms at work: The new contested terrain of control. Academy of Management Annals, 14(1), 366-410. https://doi.org/10.5465/annals.2018.0174