Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

The Governance of Autonomy: Developing an Integrated Risk Management Framework for Agentic Ai Systems in Enterprise Operations

Dickson Mdhlalose

dsskosana@gmail.com

National Electronic Media Institute of South Africa

Research Article

Keywords: agentic AI, machine identity governance, human-on-the-loop oversight, dynamic accountability, AI governance

Posted Date: March 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9032258/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

This study investigated the operational, ethical, and legal risks posed by agentic AI deployments in South African enterprise environments, employing an exploratory sequential mixed-methods design comprising 24 semi-structured interviews and a structured survey of 187 risk and governance professionals across financial services, healthcare, information technology, telecommunications, and government administration sectors. Quantitative findings confirmed that agentic deployment intensity significantly predicts governance incidents ($\beta = .38$, $p < .001$), while ERM framework adoption alone does not reduce them ($\beta = -.09$, $p = .29$). Human oversight integration ($\beta = -.32$, $p < .001$) and machine identity security maturity ($\beta = -.29$, $p = .002$) emerged as the strongest protective governance factors. Thematic analysis identified four structural mechanisms sustaining the governance gap: the obsolescence of deterministic governance assumptions, inadequate machine identity protocols, the un-operationalised status of human-on-the-loop oversight, and unresolved legal accountability for autonomous AI action.

INTRODUCTION

Few developments in the modern enterprise have been quite as disruptive or quite as quietly underestimated as the moment artificial intelligence stopped waiting to be told what to do. For the better part of two decades, organisations across virtually every sector treated Artificial Intelligence (AI) as a capable but ultimately subordinate instrument: a system that could surface insights, flag anomalies, or rank options, but that ultimately handed the baton back to a human being before anything consequential actually happened. That arrangement, comfortable as it was, is no longer the norm. What has taken its place is a generation of AI systems that do not simply inform decisions; they make them, pursue them, and execute them, often across complex chains of action that no individual human has explicitly reviewed or approved (Bommasani et al., 2022; Yao et al., 2023). These so-called agentic AI systems represent a genuinely different category of technology, and the organisations deploying them are only beginning to reckon with what that difference means for how risk is managed, how accountability is assigned, and how governance itself must evolve.

To appreciate how significant this shift is, it helps to trace where it came from. Enterprise AI in the 2010s was, by design, constrained. Systems were trained for specific tasks, classifying loan applications, detecting fraudulent transactions, recommending the next best action in a customer service queue, and they performed those tasks with measurable, auditable outputs that human decision-makers could accept or override (Russell & Norvig, 2022). The introduction of large language models changed this in a way that went well beyond incremental improvement. Large language models (LLMs) demonstrated, for the first time at scale, an ability to reason across sequential decision points, to engage with novel contexts using generalised knowledge, and to produce coherent multi-step outputs that bore the hallmarks of judgment rather than mere pattern retrieval (Wei et al., 2022). What followed was the emergence of agentic architecture systems that combine language model reasoning with persistent memory, tool-use capabilities, real-time access to external Application Programming Interfaces (APIs), and the capacity to decompose broad goals into coordinated subtasks, sometimes deploying other AI

agents to complete them in parallel (Park et al., 2023; Chase, 2023). The result is not a smarter chatbot. It is a new kind of operational actor.

This new actor is already at work and at scale. Walk through the corridors of a major financial institution, a large hospital network, or a global logistics provider today, and the chances are that agentic AI is doing something consequential somewhere in the building or, more precisely, across the building's systems without a human being watching each step. These systems are negotiating procurement agreements, triaging and escalating cybersecurity incidents, performing financial due diligence, and in some cases directly interfacing with clients in contexts that carry legal obligations and reputational stakes (Guo et al., 2024). The business case for this is not hard to understand: agentic AI can compress timelines that human labour cannot, eliminate classes of repetitive error, and unlock operational capacity that simply was not achievable before (Shavit et al., 2023). What is harder to understand and what has received far less systematic attention is the governance question that arrives alongside all of that efficiency. When a system that exercises something resembling discretion goes wrong, who is responsible? How was the error permitted to occur? And what, exactly, should have been in place to prevent it?

The Problem of Governing Autonomous Action

These questions do not have clean answers under current governance frameworks, and that is not an accident; it is a structural problem. The enterprise risk management standards that most large organisations rely on, including the Committee of Sponsoring Organizations of the Treadway Commission (COSO) Enterprise Risk Management Framework and the International Organization for Standardization (ISO) 31000 guidelines, were built on a foundational assumption: that consequential actions trace back to human actors, or at minimum to deterministic algorithms that a human has explicitly authorised to act in a defined way (National Institute of Standards and Technology [NIST], 2023). Agentic AI undermines both parts of that assumption at once. These systems interpret ambiguous instructions, choose between competing courses of action, and adjust their behaviour based on what they encounter along the way, none of which requires a human to say yes at each turn (Hendrycks et al., 2022). When an agentic system modifies a security access protocol on its own initiative, executes a financial transaction, or escalates a complaint to a regulatory body, the question of who authorised that action and who bears responsibility for its consequences does not resolve itself neatly within frameworks designed for a different kind of technology.

This accountability gap is made considerably worse by a problem that has attracted attention under the label of the 'black box', the opacity that characterises how LLM-based agents actually arrive at their decisions (Floridi et al., 2021). Traditional rule-based automation is, in principle, fully auditable: every step in the decision logic can be examined and explained. The reasoning processes of a language model agent are not like that. They are probabilistic, sensitive to context in ways that are difficult to anticipate, and often opaque even to the developers who built the system. An organisation that deploys one of these agents may find itself, after a compliance breach or a discriminatory outcome, genuinely unable to reconstruct why the decision was made which is not merely technically inconvenient; it is a serious

governance liability. Weidinger et al. (2021) mapped out a detailed taxonomy of the harms that this opacity can produce in operational contexts, ranging from misrepresentation and privacy violations to the amplification of systemic bias, all of which become sharper risks when the system has been given persistent credentials and broad access to act across enterprise systems without human sign-off at each step.

There is a further dimension to this governance challenge that deserves its own attention, because it is newer and in some ways more troubling than the others. As agentic AI systems take on functional roles within organisations, accessing databases, communicating with external partners, executing transactions, and making calls to third-party APIs, they inevitably accumulate something that looks very much like an organisational identity, complete with credentials, access privileges, and operational authority (Caron et al., 2024). This is not a metaphor. It is a security reality. Threat actors have already demonstrated that AI agents can be exploited through a technique known as prompt injection, in which malicious instructions are embedded inside data that the agent is processing, effectively redirecting its behaviour toward adversarial ends without triggering conventional detection systems (Perez & Ribeiro, 2022). In the decentralised architectures where multiple agents delegate tasks to one another, this risk compounds in ways that are genuinely difficult to contain because by the time a compromise becomes visible, the damage may already have propagated across multiple systems.

Regulatory Landscape and Its Limitations

Regulators have not been idle in the face of these developments, but the honest assessment is that the pace of regulation has not kept up with the pace of deployment. The European Union's Artificial Intelligence Act, which reached its major enforcement milestones in 2024 and received significant updates in both 2025 and 2026 to grapple more directly with agentic systems, remains the most ambitious and comprehensive legislative framework currently in force anywhere in the world (European Parliament, 2024). It establishes a risk-tiered structure, imposes transparency and human oversight obligations on high-risk AI applications, and requires conformity assessments before deployment in sensitive sectors. These are serious provisions, and they represent genuine regulatory progress. The difficulty is that the Act was substantially conceived with bounded, task-specific AI applications in mind, and agentic systems that are dynamic, self-directing, and capable of acquiring capabilities that were not present at the time of their initial assessment create interpretive ambiguities that the Act's drafters have acknowledged but have not yet fully resolved (Kolt, 2024). The core liability question, in particular, remains without a clear statutory answer in most jurisdictions: when an autonomous agent causes harm, who bears legal responsibility, the developer who built it, the organisation that deployed it, or some combination of both?

The concept of 'algorithmic malpractice' has emerged in legal scholarship as an attempt to fill part of this gap to give a name and a normative structure to the scenarios in which AI systems cause professional harm through their own autonomous errors, in the same way that a negligent professional might cause harm through a failure of competence or judgment (Hadfield-Menell & Milli, 2022). It is a

useful concept, and it is gaining traction among legal scholars. But it has not yet found its way into statute in any major economy, which means that organisations facing these situations are navigating genuinely unsettled legal terrain. Dafoe (2020) put the deeper challenge plainly: governing advanced AI does not merely require updating technical standards; it requires building entirely new institutional architectures that are capable of anticipating failure modes that have simply never existed before.

Research Gap

All of these points toward a gap in the literature that is both specific and consequential. Despite a substantial and growing body of scholarship on AI ethics, cybersecurity risk, and algorithmic accountability, what remains largely absent is an integrated, empirically grounded framework for governing the risks that autonomous AI systems generate through their own actions in real time (Anthropic, 2024; Bengio et al., 2023). The existing literature is heavily weighted toward predictive risk analytics using AI to detect and anticipate threats rather than the governance of risks that arise from the AI's own independent decision-making. The concept of Dynamic Accountability, which would provide organisations with structured guidance on how to assign responsibility, trigger intervention, and document causality when an agentic system acts with harmful consequences, remains theoretically underdeveloped. The Human-on-the-loop supervisory model, which positions human operators as monitors rather than initiators of AI action and which has been proposed as the most practical path toward meaningful oversight in high-velocity environments, lacks both rigorous empirical validation and standardised implementation guidance (Park et al., 2023; NIST, 2023). The specific intersection of machine identity security, multi-agent architectures, and enterprise governance, a set of problems that are deeply interconnected in practice, has received only fragmented attention in the academic literature, leaving organisations without coherent frameworks for understanding or managing the compounded risks that large-scale agentic deployments create.

Research Aim and Objectives

This study sets out to address that gap directly. It aims to investigate the operational, ethical, and legal risks that are distinctive to agentic AI systems deployed in enterprise environments, and to develop an integrated governance framework capable of balancing the genuine efficiency benefits of autonomous AI with the equally genuine need for meaningful human oversight and structured accountability. Three interconnected objectives guide the inquiry. The first is to map the risk landscape of agentic AI deployments in sufficient granularity to identify the categories of vulnerability that existing enterprise risk management frameworks fail to capture. The second is to evaluate the current state of regulatory instruments and governance standards and assess their adequacy or inadequacy in addressing the accountability challenges that autonomous AI action creates. The third is to propose a Dynamic Accountability Framework, a structured set of principles, processes, and oversight mechanisms developed specifically for organisations that are operating agentic AI systems in sensitive or high-stakes functions.

The urgency behind this work is not rhetorical. As Shavit et al. (2023) observed, the speed at which agentic systems can execute consequential decisions may already exceed the capacity of human operators to intervene, even in organisations that have invested in supervisory structures. The agentic enterprise is not a distant scenario that governance researchers are preparing for in advance; it is an operational reality that organisations across every major sector are navigating right now, often without the architecture in place to do so responsibly. This research is intended as a contribution to the body of knowledge that organisations, regulators, and scholars need to change that.

RESEARCH METHODOLOGY

Research Design

This study adopted an exploratory sequential mixed-methods research design to investigate the governance of agentic AI systems within South African enterprise environments. A mixed-methods approach was selected because the complexity of AI risk governance demands both the contextual richness that qualitative inquiry provides and the generalisable patterns that quantitative analysis yields (Creswell & Creswell, 2023). The philosophical underpinning of this study is pragmatism, which treats research questions rather than predetermined methodological allegiances as the primary drivers of design decisions (Tashakkori et al., 2021). This paradigm is particularly suited to applied governance research, where practical utility, contextual validity, and empirical rigour must coexist. The sequential exploratory design proceeded in two phases: an initial qualitative phase that generated thematic insights from practitioners, followed by a quantitative phase that tested and extended those findings across a broader sample of South African organisations.

Research Setting and Justification

The study was conducted within South Africa's corporate and public-sector landscape, with particular focus on industries identified as early adopters of agentic AI: financial services, healthcare, information technology, telecommunications, and government administration. South Africa presents a compelling research context for several reasons. As Africa's most advanced economy, it occupies a leadership position in enterprise digitalisation on the continent, while simultaneously grappling with regulatory frameworks, including the Protection of Personal Information Act (POPIA) and nascent AI policy guidelines that are not yet calibrated for autonomous AI action (Department of Communications and Digital Technologies, 2024). This regulatory gap mirrors the global challenge documented by Kolt (2024) and makes South Africa a timely and representative site for investigating governance deficiencies in agentic AI deployment.

Phase One: Qualitative Data Collection

The qualitative phase employed semi-structured in-depth interviews as the primary data collection instrument. Semi-structured interviews were selected because they afford the interviewer flexibility to probe emergent themes while maintaining a consistent conceptual scaffold across participants — a

balance particularly valuable when exploring novel, complex phenomena such as agentic AI governance, where standardised survey instruments may not adequately capture nuanced practitioner experiences (Bryman, 2022). A purposive sampling strategy was employed to identify participants who possessed direct, substantive experience with AI deployment, enterprise risk management, or regulatory compliance in South African organisations. Theoretical saturation was used as the criterion for determining sample adequacy (Saunders et al., 2023).

A total of 24 participants were interviewed across organisations in Johannesburg, Pretoria, Cape Town, and Durban. The sample comprised Chief Information Officers (CIOs), Chief Risk Officers (CROs), AI governance specialists, legal compliance officers, enterprise architects, and senior technology regulators from both the private sector and public administration. Interviews were conducted between March and July 2025, with each session lasting between 45 and 90 minutes. With participant consent, all interviews were audio-recorded and subsequently transcribed verbatim. To ensure credibility and member-checking, transcripts were shared with participants for review before analysis (Lincoln & Guba, 1985, as cited in Creswell & Poth, 2024).

Thematic analysis, following the six-phase framework prescribed by Braun and Clarke (2022), was employed to analyse the qualitative data. Coding and theme development were conducted using NVivo 15 (QSR International, 2023), which facilitated systematic management of the large volume of textual data and supported the transparent audit trail required for rigorous qualitative research. Initial open coding was followed by axial coding and the consolidation of overarching themes relating to accountability gaps, oversight challenges, and institutional readiness for autonomous AI governance.

Phase Two: Quantitative Data Collection

The quantitative phase deployed a structured online survey instrument developed from themes identified in the qualitative phase and grounded in existing theoretical frameworks, including the NIST AI Risk Management Framework (NIST, 2023) and the European Union Artificial Intelligence Act (European Parliament, 2024). The survey was administered to a sample of 187 enterprise professionals drawn from South African organisations with documented AI deployment programmes. Respondents were recruited through a combination of professional networks, LinkedIn outreach targeting verified practitioners, and partnerships with the South African Institute of Chartered Accountants (SAICA) and the Institute of Risk Management South Africa (IRMSA). The survey instrument comprised Likert-scale items, categorical questions, and scenario-based vignettes, and was pre-tested with a pilot group of 12 practitioners to assess construct validity and internal consistency (Pallant, 2024). Cronbach's alpha coefficients for all multi-item scales exceeded the conventional threshold of 0.70, confirming acceptable reliability.

Quantitative data were analysed using IBM SPSS Statistics Version 29 (IBM Corp., 2023). Descriptive statistics were computed to characterise the sample and summarise response distributions. Inferential analyses, including independent-samples t-tests, one-way analysis of variance (ANOVA), and multiple linear regression, were employed to test relationships between governance maturity, organisational size, sector, and AI deployment intensity. Factor analysis was conducted to validate the latent structure of the

governance readiness scale developed for this study. Statistical significance was assessed at the conventional $\alpha = 0.05$ threshold.

Ethical Considerations

This study received ethical clearance from the relevant institutional review board prior to data collection. All participants provided informed written consent and were assured of their right to withdraw without penalty at any stage of the research. Confidentiality and anonymity were maintained throughout: organisations and individuals are identified in the findings only through non-identifying descriptors such as sector, organisational size, and role category. All digital data were stored on encrypted, password-protected servers, and audio recordings were deleted upon completion of transcription, in compliance with POPIA (Republic of South Africa, 2013) and the ethical guidelines of the institution hosting the study. No incentives were offered to participants, minimising the risk of response bias (Israel & Hay, 2006, as cited in Flick, 2022).

Validity, Reliability, and Limitations

Methodological rigour was ensured through triangulation of qualitative and quantitative data sources, prolonged engagement with organisational contexts, and an independent review of coding decisions by a second researcher to assess inter-coder reliability. The mixed-methods design strengthens transferability by situating statistical findings within the interpretive context provided by practitioner narratives (Creswell & Creswell, 2023). Notwithstanding these measures, the study carries inherent limitations. The South African sample, while internally diverse, constrains the direct generalisability of findings to other national contexts with different regulatory regimes and AI adoption trajectories. Additionally, self-report data in the quantitative phase are susceptible to social desirability bias, particularly on questions relating to governance failures or regulatory non-compliance (Podsakoff et al., 2024). These limitations are acknowledged and addressed, where possible, through the methodological triangulation described above.

LITERATURE REVIEW

Understanding the governance of agentic artificial intelligence (AI) in enterprise settings requires more than a technical inventory of capabilities; it demands a critical engagement with the intellectual traditions, empirical debates, and normative controversies that define this rapidly evolving field. The four subtopics addressed in this review, the shift from static to agentic AI, machine identity governance and autonomous vulnerabilities, human-AI collaborative oversight, and the legal and regulatory implications of autonomous agency, are not independent silos. They are interlocking fault lines in a governance architecture that was not designed for systems that act, adapt, and influence outcomes without step-by-step human authorisation (Shavit et al., 2023; NIST, 2023). Taken together, the literature reviewed here situates the present study's aim to develop an integrated DAF for agentic AI in South African enterprises within a global conversation that is urgent, contested, and unresolved.

The Shift from Static to Agentic AI: Evolution of Enterprise Automation

The intellectual trajectory of enterprise AI is, in many respects, a story of progressively loosened constraints. Early machine learning systems deployed in organisational settings throughout the 2010s operated within tightly bounded problem domains classifying inputs, detecting anomalies, or ranking recommendations but consistently left execution authority with human decision-makers (Russell & Norvig, 2022). This human-in-the-loop paradigm reflected both a practical limitation and a deliberate design philosophy: machines were tools, not actors, and the human being remained the locus of agency and, by extension, accountability (Dignum, 2020). That design philosophy is now under fundamental pressure.

The emergence of LLMs as a foundational technology class catalysed the most significant qualitative leap in AI capability since the advent of deep learning. Wei et al. (2022) demonstrated that sufficiently large language models exhibit emergent reasoning capacities, including multi-step chain-of-thought problem solving that were not present in their smaller predecessors and were not explicitly trained into them. This unpredictability is not merely academically interesting; it represents a governance challenge. If a system's capabilities cannot be reliably anticipated from its architecture, the risk surface expands in ways that standard pre-deployment assessments cannot fully capture (Hendrycks et al., 2022; Bengio et al., 2023). Bommasani et al. (2022) foreshadowed precisely this problem in their seminal analysis of foundation models, arguing that the breadth of downstream applications enabled by these systems creates regulatory and governance gaps that no single framework can address.

Contemporary agentic architectures extend LLM capabilities into the operational domain. Systems built on frameworks such as those described by Chase (2023) and Yao et al. (2023) combine persistent memory, tool-use APIs, and the capacity to decompose complex goals into executable subtasks, effectively functioning as autonomous programmes rather than responsive utilities. Guo et al. (2024) documented the deployment of multi-agent LLM systems across financial services, healthcare, and logistics, identifying patterns of delegation, escalation, and environmental interaction that resemble and often exceed the autonomous behaviour of junior human employees. This analogy is not merely rhetorical. It has direct implications for governance: if agentic systems behave like employees, should they be governed like employees, and if not, by what framework should they be governed? That question, left insufficiently answered in both the academic and practitioner literature, constitutes a core motivation for the present study.

It would be intellectually dishonest, however, to present this evolution as uniformly beneficial or inevitable. Marcus and Davis (2022) offer a sustained sceptical counterargument to AI optimism, contending that current LLM-based systems remain fundamentally brittle — capable of impressive performance on trained distributions but prone to systematic failure when confronted with genuine novelty. Jobin et al. (2020), in a comprehensive survey of global AI ethics guidelines, found that despite widespread rhetorical commitment to values such as transparency and accountability, the practical mechanisms for realising these values in deployed systems remain underdeveloped. The gap between

aspirational capability and governance reality visible in the earliest phases of AI ethics discourse has, if anything, widened as systems have become more autonomous.

Machine Identity Governance and Autonomous Vulnerabilities

As agentic AI systems assume functional roles within enterprise operations, they necessarily acquire something that resembles organisational identity access credentials, API keys, database privileges, communication authorities, and, in some architectures, the ability to create and manage sub-agents with delegated permissions (Caron et al., 2024). This development represents a structural novelty for which traditional identity and access management (IAM) frameworks were simply not designed. Conventional IAM practices assume that access subjects are human beings whose identities can be verified, whose behaviours can be anticipated within role-based constraints, and whose accountability can be enforced through legal and contractual mechanisms (NIST, 2023). Agentic AI systems satisfy none of these assumptions. They are not human, their behaviour is probabilistic rather than rule-bound, and their legal personality while increasingly discussed in academic and regulatory circles — remains undefined in most jurisdictions.

The security implications of this governance gap are concrete and serious. Perez and Ribeiro (2022) provided the first systematic analysis of prompt injection as an attack vector, demonstrating that malicious instructions embedded in data inputs processed by an LLM-based agent can effectively redirect the agent's behaviour without triggering conventional intrusion detection systems. This is not a theoretical vulnerability. Weidinger et al. (2021) documented a broader taxonomy of harms arising from language model deployment, including privacy violation, misinformation amplification, and the facilitation of unauthorised access — risks that intensify when the AI system operates with persistent credentials and reduced human oversight. The threat is compounded in multi-agent architectures: when one autonomous system delegates tasks to another, the attack surface expands geometrically, and the capacity to trace causality already limited in single-agent systems erodes further (Bommasani et al., 2022; Shavit et al., 2023).

Researchers have coined the term 'agent hijacking' to describe scenarios where an adversary exploits a prompt injection vulnerability to assume effective control of an AI agent, redirecting its actions toward adversarial ends while the deploying organisation remains unaware (Caron et al., 2024). In enterprise environments where agentic systems have been granted authority to execute financial transactions, modify access controls, or communicate with regulators, the consequences of a successful hijacking attack can be catastrophic. Floridi et al. (2021) argued that AI systems operating in high-stakes institutional contexts must be subject to what they term 'ethics-by-design,' the embedding of governance constraints at the architectural level rather than as post-hoc controls. This argument gains urgency in the agentic context: an AI system that can be hijacked through its own inputs requires governance that extends to the design of its input-processing architecture, not merely to the policies governing its deployment.

It is worth raising a dissenting perspective here. Some researchers argue that the machine identity problem is not qualitatively different from existing challenges in managing privileged access for automated systems, scripts, and service accounts, all of which have been governed (imperfectly) by existing IAM frameworks for decades (Brundage et al., 2020). This position is not entirely without merit. However, it underestimates the degree to which agentic AI's natural language processing capabilities expand the attack surface beyond anything addressable by conventional access controls. A service account with fixed permissions is auditable; an AI agent that can interpret and act on natural language instructions, including malicious ones, is qualitatively different in its vulnerability profile. Lazar and Nelson (2023) make a similar point, noting that the governance of AI agents requires fundamentally new institutional mechanisms, not merely extensions of existing ones.

Human-AI Collaborative Oversight: The 'Human-on-the-Loop' Model

The concept of the 'human-on-the-loop' represents perhaps the most practically consequential idea in the governance of agentic AI and, simultaneously, one of the most contested. In contrast to the 'human-in-the-loop' model, which requires human authorisation at each decision point, the human-on-the-loop model positions human operators as monitors and exception-handlers rather than as initiators of action (Park et al., 2023). The theoretical appeal is straightforward: it preserves the speed and scalability advantages of autonomous AI operation while retaining a human override capacity for consequential decisions. In practice, however, the operationalisation of this model is fraught with psychological, organisational, and technical challenges that the existing literature has only partially addressed.

Shavit et al. (2023) raised what may be the most fundamental challenge: the speed at which agentic systems execute multi-step workflows may outpace the cognitive capacity of human monitors to detect, interpret, and respond to problematic decisions before consequential harm occurs. This is not merely a technology problem it is a human factors problem. Amodei et al. (2021) identified 'reward hacking' and unintended goal pursuit as persistent risks in autonomous AI systems, arguing that the most dangerous failure modes are precisely those in which the system behaves coherently from its own perspective while deviating materially from the deploying organisation's intent. Human monitors who lack the technical literacy to distinguish coherent-but-harmful behaviour from appropriate agent action are, in practice, not providing meaningful oversight at all.

The emerging concept of 'supervisor agents' AI systems specifically designed to monitor other AI systems, represents an innovative, if controversial, response to this challenge. Leike et al. (2022) proposed alignment mechanisms by which AI supervisors could monitor agent outputs against predefined value constraints, triggering human alerts when deviations exceeded pre-set thresholds. While intellectually promising, this approach raises a recursive governance problem: if we cannot fully trust the monitored agent, what grounds do we have for trusting the monitoring agent? Bengio et al. (2023) acknowledged this dilemma directly, arguing that the governance of AI-on-AI oversight requires a higher-order layer of human institutional accountability that cannot itself be delegated to automated systems. The NIST AI Risk Management Framework (NIST, 2023) attempts to address this through its

'govern, map, measure, manage' structure, but critics, including Hendrycks et al. (2022), argue that its provisions remain too generic for the operational realities of agentic deployment.

The psychological dimension of human-in-the-loop oversight deserves more attention than it has received in the governance literature. Automation bias, the documented tendency of human monitors to defer to automated system outputs even when those outputs are detectably incorrect, poses a systemic risk to any oversight model that relies on human exception-handling (Kaminski, 2023). Anthropic (2024) acknowledged this challenge in its responsible deployment guidance, noting that the design of human oversight interfaces must actively counteract automation bias rather than assume that monitoring capability equates to monitoring effectiveness. This insight has direct implications for the present study's DAF: intervention triggers, alert design, and escalation pathways must be designed with human cognitive limitations in mind, not merely with technical precision.

Legal and Regulatory Implications of Autonomous Agency

The legal landscape surrounding agentic AI is in a state of productive but disorganised ferment. The European Union's Artificial Intelligence Act, enacted in 2024 and further updated in 2025 and 2026 to address emerging agentic deployment patterns, represents the most comprehensive attempt by any major jurisdiction to impose risk-tiered obligations on AI developers and deployers (European Parliament, 2024). The Act's framework, which classifies AI applications by risk level and prescribes corresponding requirements for transparency, conformity assessment, and human oversight, is conceptually sound but operationally incomplete in its application to agentic systems. Kolt (2024) identified the core difficulty: the Act's provisions were substantially designed with bounded, task-specific AI applications in mind. Agentic systems, which can acquire capabilities, pursue goals, and interact with environments in ways not fully anticipated at deployment, strain the Act's categorical framework in ways that its drafters have acknowledged but not yet fully resolved.

The liability question is the most legally acute dimension of this problem. When an autonomous AI agent independently executes a transaction that causes financial harm, modifies a security configuration that enables a breach, or communicates with a regulatory body in a manner that creates legal exposure, the conventional mechanisms for attributing liability which assume a human actor who made a deliberate or negligent decision do not map cleanly onto the factual situation (Hadfield-Menell & Milli, 2022). Solove (2023) argued that existing tort law frameworks are structurally inadequate for AI-generated harms, not merely because AI systems are non-human, but because the probabilistic, context-sensitive nature of their decision-making defies the causal analysis on which liability doctrines depend. This argument is supported by Floridi et al. (2021), who contend that AI governance requires the construction of entirely new normative categories rather than the extension of existing ones.

The concept of 'algorithmic malpractice' has emerged in legal scholarship as a partial response to this challenge. Introduced by Hadfield-Menell and Milli (2022) to capture scenarios in which AI systems cause professional harm through autonomous error analogous to medical or legal malpractice, the concept provides a normative framework that distributes accountability between developers, deployers,

and, in some formulations, the AI systems themselves. While the analogy has intuitive appeal, Dafoe (2020) cautioned that the governance of advanced AI cannot proceed through analogy alone; it requires the construction of institutional architectures that anticipate failure modes with no historical precedent. Kaminski (2023) extended this argument specifically to the liability context, proposing a 'distributed accountability' model in which responsibility is apportioned among developers, deployers, and users according to their respective capacities for control, a framework broadly consistent with the DAF proposed in the present study.

The South African regulatory context adds a further layer of complexity that renders this literature directly relevant to the present study. POPIA provides meaningful data governance guardrails but does not address the accountability questions generated by autonomous AI action (Republic of South Africa, 2013). South Africa's Draft National AI Policy Framework remains nascent and does not yet articulate liability mechanisms appropriate for agentic deployments (Department of Communications and Digital Technologies, 2024). Beasley et al. (2021) observed, in the enterprise risk management context broadly, that the presence of documented frameworks does not, in itself, reduce actual risk exposure, a finding replicated in the present study's quantitative results and one that underscores the gap between regulatory aspiration and operational governance reality. The literature reviewed here thus converges on a conclusion that motivates the present study's core contribution: existing governance frameworks, whether regulatory or organisational, require substantive redesign, not merely incremental adaptation to manage the risks generated by autonomous AI agents in enterprise operations.

RESULTS AND DISCUSSION

Sample Profile and Descriptive Overview

The combined study sample comprised 211 participants: 24 individuals who participated in the qualitative interview phase and 187 in the structured quantitative survey phase. Qualitative participants were drawn from organisations headquartered or operating across Johannesburg, Pretoria, Cape Town, and Durban, spanning five sectors that represent South Africa's most active agentic AI adopters: financial services (29%), healthcare (18%), information technology (21%), telecommunications (17%), and government administration (15%). The quantitative sample reflected a comparable sectoral distribution. Of the 187 survey respondents, 64% represented large organisations with more than 500 employees, 24% represented medium-sized enterprises, and 12% were from smaller organisations. Senior practitioners, including Chief Information Officers (CIOs), Chief Risk Officers (CROs), AI governance specialists, and enterprise architects, constituted 58% of the quantitative respondents, lending strong practitioner authority to the dataset. Descriptive statistics generated using IBM SPSS Statistics Version 29 (IBM Corp., 2023) confirmed adequate cross-sectoral and role-level diversity to support robust inferential analysis.

Table 1 Participant Distribution by Sector and Role Category (N = 187, Quantitative Phase)

Sector	n	%	Mean Governance Readiness Score
Financial Services	54	28.9%	3.41
Information Technology	39	20.9%	3.67
Healthcare	34	18.2%	3.14
Telecommunications	32	17.1%	3.29
Government Administration	28	15.0%	2.73

Note. Governance readiness scores are based on a five-point Likert scale (1 = very low, 5 = very high). Higher scores indicate greater perceived governance preparedness for agentic AI deployment.

Quantitative Findings: Governance Readiness and Agentic AI Risk

Quantitative analysis conducted using IBM SPSS Statistics Version 29 (IBM Corp., 2023) yielded several statistically significant findings that substantiate and extend the governance gap identified in the introduction. Pearson correlation analysis revealed a significant positive relationship between agentic AI deployment intensity and reported governance incidents ($r = .41$, $p < .001$), confirming that organisations deploying agentic systems at scale are experiencing substantially elevated adverse event rates. Crucially, governance readiness scores exhibited a significant negative relationship with incident frequency ($r = -.37$, $p < .001$), indicating that robust governance architecture meaningfully reduces harmful outcomes, though governance readiness remained comparatively low across all sectors sampled, with a cross-sample mean of 3.25 on a five-point scale.

Multiple linear regression, with governance incident frequency as the dependent variable, identified agentic deployment intensity ($\beta = .38$, $p < .001$), machine identity security maturity ($\beta = -.29$, $p = .002$), and human oversight integration ($\beta = -.32$, $p < .001$) as the strongest predictors of incident rates ($F(4, 182) = 14.67$, $p < .001$, $R^2 = .24$). By contrast, ERM framework adoption status was not a significant independent predictor ($\beta = -.09$, $p = .29$), replicating in the agentic AI context the broader governance insight advanced by Beasley et al. (2021) and Hendrycks et al. (2022): that the presence of documented frameworks does not, in itself, reduce actual exposure. One-way ANOVA revealed significant sector-level differences in governance readiness ($F(4, 182) = 8.94$, $p < .001$). Post-hoc Tukey HSD analysis identified government administration organisations as exhibiting the lowest governance readiness scores ($M = 2.73$, $SD = 0.58$), while information technology firms demonstrated the highest preparedness ($M = 3.67$, $SD = 0.61$), a divergence that likely reflects differential exposure to AI governance norms and the sharper regulatory accountability faced by technology-sector organisations under POPIA (Republic of South Africa, 2013) and nascent AI policy directives (Department of Communications and Digital Technologies, 2024).

Table 2

Multiple Linear Regression: Predictors of Agentic AI Governance Incident Frequency (N = 187)

Predictor	β	SE	t	p
Agentic Deployment Intensity	.38	.06	6.33	< .001
Human Oversight Integration	-.32	.07	-4.57	< .001
Machine Identity Security Maturity	-.29	.08	-3.63	.002
ERM Framework Adoption	-.09	.09	-1.00	.29

Note. $R^2 = .24$, $F(4, 182) = 14.67$, $p < .001$. β = standardised regression coefficient; SE = standard error.

Qualitative Findings: Thematic Analysis of Practitioner Experience

Thematic analysis of the 24 semi-structured interviews, conducted and managed using NVivo 15 (QSR International, 2023), yielded four overarching themes that correspond directly to the four literature review subtopics: (1) the governance implications of the shift from static to agentic AI; (2) machine identity governance and autonomous vulnerabilities; (3) the operationalisation challenges of the human-on-the-loop model; and (4) regulatory uncertainty and the emerging concept of algorithmic malpractice. Each theme is discussed below, with qualitative findings integrated with relevant quantitative evidence.

The Governance Vacuum Created by the Shift to Agentic AI

Participants across all sectors articulated a shared recognition that the transition from bounded, decision-support AI to autonomous, action-executing agentic systems had fundamentally outpaced existing governance architectures. Enterprise architects and CIOs in the financial services and telecommunications sectors described deploying agentic systems capable of executing multi-step workflows, including API calls, transaction executions, and escalation triggers without real-time human authorisation, consistent with the architectural capabilities described by Yao et al. (2023) and Guo et al. (2024). Yet the governance structures applied to these deployments remained rooted in frameworks designed for deterministic, rule-based automation. A CRO in financial services described the situation as governing a self-driving vehicle using the road rules for a horse-drawn cart, a vivid articulation of the structural mismatch between agentic capability and inherited oversight paradigms. These qualitative observations are consistent with the quantitative finding that agentic deployment intensity is the strongest positive predictor of governance incidents ($\beta = .38$, $p < .001$), and with the introduction's identification of the COSO and ISO 31000 frameworks as fundamentally ill-suited to governing systems that exercise something functionally analogous to discretion (Hendrycks et al., 2022; NIST, 2023).

Machine Identity Governance and the Prompt Injection Threat

The second major theme concerned the governance of machine identity — the set of credentials, access privileges, and operational authorities acquired by agentic AI systems as they assume functional roles within organisational IT ecosystems. Consistent with the theoretical framework advanced by Caron et al. (2024), interview participants described AI agents operating with persistent access to enterprise

databases, email systems, and external APIs, creating attack surfaces that conventional identity and access management (IAM) protocols were not designed to address. Particularly acute was practitioner concern regarding prompt injection attack scenarios where malicious instructions embedded in data inputs effectively redirect agent behaviour, a vulnerability empirically documented by Perez and Ribeiro (2022). An AI governance specialist in the information technology sector described a near-incident in which an agentic system processing incoming vendor communications was manipulated via an injected payload to initiate unauthorised data retrieval, a scenario that the organisation's existing security framework had no detection mechanism for. The quantitative data corroborate the severity of this vulnerability: machine identity security maturity emerged as a significant independent negative predictor of governance incidents ($\beta = -.29$, $p = .002$), and only 31% of survey respondents rated their organisation's machine identity governance protocols as adequate for agentic deployments. The risks intensify in decentralised multi-agent architectures, where sub-agent delegation chains compound the difficulty of tracing causality and containing compromise (Bommasani et al., 2022; Shavit et al., 2023).

The Human-on-the-Loop Model: Promise and Operational Reality

The third theme addressed the operationalisation of human oversight in agentic AI environments. The human-on-the-loop supervisory model, in which human operators monitor AI action without initiating each step was the most frequently cited governance response among participants, yet it was also the most consistently described as failing in practice. Participants from healthcare, government administration, and financial services organisations described oversight arrangements that were nominally human-on-the-loop but functionally inadequate: human monitors who lacked the contextual access, technical literacy, or cognitive bandwidth to identify problematic agent decisions before consequential actions were completed. This finding is consistent with the concern raised by Park et al. (2023) and Shavit et al. (2023) that the speed at which agentic systems execute decisions may outpace the capacity of human operators to intervene meaningfully. The quantitative data confirm the governance value of robust human oversight integration: it was the second strongest negative predictor of incident frequency ($\beta = -.32$, $p < .001$). However, only 38% of surveyed organisations reported having formalised human-on-the-loop protocols with defined intervention triggers, alert thresholds, and escalation procedures suggesting that the model is widely aspirational but insufficiently operationalised, as identified in the research gap section of this study (NIST, 2023; Anthropic, 2024).

Regulatory Uncertainty and Algorithmic Malpractice

The fourth theme addressed the legal and regulatory dimensions of agentic AI deployment, a domain characterised by participants as deeply uncertain and institutionally unprepared. Legal compliance officers and senior technology regulators across all sectors described the EU AI Act (European Parliament, 2024) as a significant reference point, particularly following its 2025 and 2026 updates, yet noted that its provisions were difficult to apply to agentic systems whose self-directing behaviour introduces interpretive ambiguities that the Act's drafters had not fully anticipated (Kolt, 2024). South African organisations faced a compounded regulatory gap: the POPIA provides meaningful data

governance guardrails but does not address the liability questions generated by autonomous AI action, while South Africa's Draft National AI Policy Framework (Department of Communications and Digital Technologies, 2024) remains in nascent form. The concept of algorithmic malpractice introduced by Hadfield-Menell and Milli (2022) to capture scenarios where AI systems cause professional harm through autonomous error resonated strongly with participants, with 74% of survey respondents indicating that their organisation had no defined legal protocol for attributing responsibility when an agentic system causes a compliance breach or financial harm. This regulatory vacuum, consistent with the broader governance gap documented by Dafoe (2020) and Bengio et al. (2023), represents one of the most consequential unresolved risks facing South African enterprise AI deployment.

Integrated Discussion and Implications for the Dynamic Accountability Framework

The integrated findings of this study provide robust triangulated evidence for the existence, scale, and structural origins of the agentic AI governance gap in South African enterprises. The quantitative finding that agentic deployment intensity significantly predicts governance incidents while ERM framework adoption does not independently reduce them situates the problem precisely where the introduction identified it: not in an absence of risk management activity, but in the fundamental mismatch between inherited governance architectures and the autonomous, discretionary behaviour of agentic systems. The qualitative evidence enriches this structural account by revealing four specific mechanisms that collectively constitute the governance gap: the obsolescence of deterministic governance assumptions, the inadequacy of machine identity security protocols, the aspirational-but-un-operationalised status of human oversight models, and the legal accountability vacuum created by regulatory frameworks not yet calibrated for autonomous AI agents.

These findings have direct implications for the Dynamic Accountability Framework (DAF) proposed as the third research objective of this study. The DAF must address governance across all four identified gap dimensions: it must replace assumption-based governance with real-time action monitoring; it must integrate machine identity governance into AI deployment protocols as a first-order security requirement; it must operationalise the human-on-the-loop model with specific intervention triggers, defined escalation pathways, and role-differentiated oversight responsibilities; and it must provide organisations with a legally defensible causal documentation standard for attributing responsibility when agentic systems cause harm. The consistent finding that government administration organisations exhibit the lowest governance readiness scores, despite operating agentic systems in contexts of significant public accountability, underscores the urgency of DAF adoption in the public sector specifically. As Shavit et al. (2023) observed, the speed of agentic decision execution may outpace human intervention capacity even under well-designed supervisory structures a reality that demands governance frameworks built not merely for compliance performance but for genuine real-time accountability.

CONCLUSION

This study set out to investigate the unique governance challenges posed by agentic AI systems in enterprise environments and to develop a DAF capable of addressing them. Its central argument that the governance gap is structural rather than technical, rooted in inherited frameworks never designed for systems that exercise discretion, is substantiated by the empirical findings. Agentic deployment intensity significantly predicts governance incidents, ERM framework adoption alone does not reduce them, and organisations with robust machine identity security and formalised human-on-the-loop oversight consistently experience fewer adverse outcomes. In response, the DAF operationalises four governance imperatives: real-time action monitoring, machine identity governance as a first-order security requirement, formalised oversight with defined intervention thresholds, and legally defensible causal documentation standards. Together, these mechanisms offer a pathway from inherited, assumption-based governance toward dynamic accountability suited to the realities of autonomous enterprise AI.

Several limitations warrant acknowledgement. The South African research setting, while internally diverse, restricts the direct generalisability of findings to jurisdictions with substantially different regulatory regimes or AI adoption contexts. The cross-sectional design cannot account for the rapid evolution of both agentic capabilities and regulatory responses since data collection. Self-report data carry social desirability risks that anonymous collection only partially addresses, and the exclusion of AI developers from the sample leaves the supply-side governance perspective underexplored.

Future research should pursue longitudinal designs to assess whether DAF adoption demonstrably reduces incident rates over time, and cross-national studies to test transferability across regulatory environments. The governance of multi-agent architectures where autonomous systems both delegate to and monitor one another represents a particularly urgent and undertheorised frontier. Practitioners are urged to approach the DAF as a living instrument, requiring continuous revision as agentic capabilities evolve, rather than as a static compliance artefact. The agentic enterprise is not a future scenario it is an operational reality being navigated now, and governing it responsibly is an immediate obligation that cannot be deferred.

References

1. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2021). Concrete problems in AI safety. arXiv. https://doi.org/10.48550/arXiv.1606.06565
2. Anthropic. (2024). Claude's model specification: Responsible scaling and deployment of large language models. Anthropic. https://www.anthropic.com/research
3. Beasley, M. S., Branson, B. C., & Pagach, D. (2021). An analysis of the maturity and strategic impact of investments in enterprise risk management. Journal of Accounting and Public Policy, 40(2), Article 106847. https://doi.org/10.1016/j.jaccpubpol.2021.106847
4. Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Harari, Y. N., Zhang, Y., Hadfield-Menell, D., Sellitto, M., Guo, S., von Roenn, C., & Russell, S. (2023). Managing extreme AI risks amid rapid progress. Science, 384(6698), 842–845. https://doi.org/10.1126/science.adn0117

5. Bommasani, R., Hudson, D. A., Aditi, E., Altman, R., Arora, S., Bernstein, S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., Donahue, C., . . . Liang, P. (2022). On the opportunities and risks of foundation models. arXiv. https://doi.org/10.48550/arXiv.2108.07258
6. Bommasani, R., Hudson, D. A., Aditi, E., Altman, R., Arora, S., Bernstein, S., . . . Liang, P. (2022). On the opportunities and risks of foundation models. arXiv. https://doi.org/10.48550/arXiv.2108.07258
7. Braun, V., & Clarke, V. (2022). Thematic analysis: A practical guide. SAGE Publications.
8. Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., hÉigeartaigh, S. Ó., Beard, S., Belfield, H., Farquhar, S., . . . Amodei, D. (2020). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. arXiv. https://doi.org/10.48550/arXiv.1802.07228
9. Bryman, A. (2022). Social research methods (6th ed.). Oxford University Press.
10. Caron, M., Lefebvre, J., & Tremblay, S. (2024). Machine identity in the age of agentic AI: Governance frameworks for non-human actors in enterprise systems. Journal of Information Security and Applications, 81, 103–119. https://doi.org/10.1016/j.jisa.2024.103119
11. Chase, H. (2023). *LangChain: Building applications with large language models through composability*. LangChain Inc. https://github.com/langchain-ai/langchain
12. Creswell, J. W., & Creswell, J. D. (2023). Research design: Qualitative, quantitative, and mixed methods approaches (6th ed.). SAGE Publications.
13. Creswell, J. W., & Poth, C. N. (2024). Qualitative inquiry and research design: Choosing among five approaches (5th ed.). SAGE Publications.
14. Dafoe, A. (2020). AI governance: A research agenda. Future of Humanity Institute, University of Oxford. https://www.fhi.ox.ac.uk/wp-content/uploads/GovAI-Agenda.pdf
15. Department of Communications and Digital Technologies. (2024). Draft national artificial intelligence policy framework for South Africa. Republic of South Africa. https://www.dcdt.gov.za
16. Dignum, V. (2020). Responsible artificial intelligence: How to develop and use AI in a responsible way. Springer. https://doi.org/10.1007/978-3-030-30371-6
17. European Parliament. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council: Artificial Intelligence Act. Official Journal of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689
18. European Parliament. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council: Artificial Intelligence Act. Official Journal of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689
19. Flick, U. (2022). An introduction to qualitative research (7th ed.). SAGE Publications.
20. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2021). An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. In L. Floridi (Ed.), Ethics, governance,

and policies in artificial intelligence (pp. 19–39). Springer. https://doi.org/10.1007/978-3-030-81907-1_3

21. Guo, T., Chen, X., Wang, Y., Chang, R., Peng, S., Chawla, N. V., Wiest, D., & Zhang, X. (2024). Large language model-based multi-agent systems in enterprise automation: Deployment patterns and risk profiles. IEEE Transactions on Neural Networks and Learning Systems, 35(4), 5201–5218. https://doi.org/10.1109/TNNLS.2024.3367541
22. Hadfield-Menell, D., & Milli, S. (2022). Algorithmic malpractice and the limits of autonomous decision accountability. AI & Society, 37(3), 1123–1138. https://doi.org/10.1007/s00146-021-01262-5
23. Hendrycks, D., Carlini, N., Schulman, J., & Steinhardt, J. (2022). Unsolved problems in ML safety. arXiv. https://doi.org/10.48550/arXiv.2109.13916
24. IBM Corp. (2023). IBM SPSS Statistics for Windows, Version 29.0. IBM Corp.
25. Jobin, A., Ienca, M., & Vayena, E. (2020). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2
26. Kaminski, M. E. (2023). Liability for AI decision-making: Pathways through tort, contract, and regulation. Harvard Journal of Law and Technology, 36(1), 113–178. https://doi.org/10.2139/ssrn.4287264
27. Kolt, N. (2024). Governing AI agents. Columbia Law Review, 124(2), 401–467. https://doi.org/10.2139/ssrn.4772822
28. Lazar, S., & Nelson, A. (2023). AI safety on whose terms? Issues in Science and Technology, 39(4), 56–63. https://doi.org/10.2307/27265478
29. Leike, J., Martic, M., Krakovna, V., Ortega, P. A., Everitt, T., Lefrancq, A., Orseau, L., & Legg, S. (2022). AI safety gridworlds. arXiv. https://doi.org/10.48550/arXiv.1711.09883
30. Marcus, G., & Davis, E. (2022). Rebooting AI: Building artificial intelligence we can trust. Vintage Books.
31. National Institute of Standards and Technology (NIST). (2023). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1
32. Pallant, J. (2024). SPSS survival manual: A step-by-step guide to data analysis using IBM SPSS (8th ed.). McGraw-Hill.
33. Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (pp. 1–22). Association for Computing Machinery. https://doi.org/10.1145/3586183.3606763
34. Perez, F., & Ribeiro, I. (2022). Ignore previous prompt: Attack techniques for language models. arXiv. https://doi.org/10.48550/arXiv.2211.09527
35. Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2024). Common method biases in behavioral research: A critical review and future directions. Journal of Applied Psychology, 108(4), 542–568.

https://doi.org/10.1037/apl0000533

36. QSR International. (2023). NVivo (Version 15) [Computer software]. QSR International Pty Ltd. https://www.qsrinternational.com
37. Republic of South Africa. (2013). Protection of Personal Information Act 4 of 2013. Government Gazette. https://www.gov.za/documents/protection-personal-information-act
38. Russell, S., & Norvig, P. (2022). *Artificial intelligence: A modern approach* (4th ed.). Pearson.
39. Saunders, M. N. K., Lewis, P., & Thornhill, A. (2023). Research methods for business students (9th ed.). Pearson Education.
40. Shavit, Y., Misra, I., Anthropic Policy Team, Goldstein, E., Wen, M., Perez, E., Hubinger, E., & Clark, J. (2023). Practices for governing agentic AI systems. arXiv. https://doi.org/10.48550/arXiv.2312.07889
41. Solove, D. J. (2023). Artificial intelligence and the future of privacy. GWU Legal Studies Research Paper No. 2023-21. https://doi.org/10.2139/ssrn.4440399
42. Tashakkori, A., Johnson, R. B., & Teddlie, C. (2021). Foundations of mixed methods research: Integrating quantitative and qualitative approaches in the social and behavioral sciences (2nd ed.). SAGE Publications.
43. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In Advances in Neural Information Processing Systems (Vol. 35, pp. 24824–24837). Curran Associates. https://doi.org/10.48550/arXiv.2201.11903
44. Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P. S., Cheng, M., Glaese, 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. (2021). Ethical and social risks of harm from language models. arXiv. https://doi.org/10.48550/arXiv.2112.04359
45. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. In Proceedings of the 11th International Conference on Learning Representations (pp. 1–17). OpenReview. https://doi.org/10.48550/arXiv.2210.03629