Moral Theories and the Ethics of Artificial Intelligence in Business

Qualifying Exam Question 1: Dr. Robert Sharp

Christopher Law

11/1/2025

Muskingum University

Department of Organizational Leadership

As firms accelerate deployment of generative artificial intelligence to gain a competitive advantage, leaders confront unsettled ethical questions that carry strategic, legal, and reputational risk. This paper employs the three major moral theories in contemporary ethics, namely deontology, utilitarianism or consequentialism, and virtue ethics, to establish a common framework for analysis, to apply that framework to concrete executive decisions in hiring, performance review, stakeholder communication, and marketing, and to synthesize recent empirical and policy evidence from 2019 to 2025. Deontology clarifies rights, duties, and constraints relevant to privacy, consent, non-deception, and explanation. Consequentialism provides a structured cost–benefit and externality analysis across the entire system lifecycle. Virtue ethics centers on character, practical wisdom, and culture, reminding leaders that technological practices form organizational habits over time.

The discussion integrates governance instruments and practices, including the EU Artificial Intelligence Act, the ISO/IEC 42001 AI management system standard, the NIST AI Risk Management Framework, as well as documentation and audits. In addition to identifying constraints, the analysis also surfaces pro-AI arguments within each tradition: deontology can endorse systems that advance duties of fairness, beneficence, and safety so long as they honor informed consent, privacy, and meaningful explanation, and consequentialism supports deployments whose net welfare gains for affected stakeholders are demonstrated through rigorous evidence and monitoring across the lifecycle. From a virtue-ethical perspective, adoption is favored when AI enhances organizational character by fostering transparency, accountability, and practical wisdom in leaders and teams. The paper concludes with a theory-informed decision approach and a set of core competencies that enable chief executives to

capture value from artificial intelligence without undermining moral legitimacy.

Deontology is a rule-based approach to moral reasoning in which an action is judged as right or wrong by reference to duties, rights, and principles rather than by reference to its outcomes. In Kant's formulation of the Categorical Imperative, one must act only on maxims that can be willed as universal law and must always treat persons as ends in themselves rather than as mere means (Kant, 1785/2011). Later deontologists, such as W. D. Ross, recognized a set of prima facie duties, including fidelity, reparation, justice, beneficence, non-maleficence, and self-improvement, and explained moral deliberation as the weighing of these duties in context (Ross, 1930/2002). On a deontological view, the ethics of a business practice, including the deployment of artificial intelligence, turn on whether it respects rights and fulfills duties, not only on whether it produces advantageous consequences.

Utilitarianism, a form of consequentialism, evaluates actions by their outcomes for overall welfare. Mill defines the moral point as the greatest happiness for the greatest number and insists on impartial consideration of interests across stakeholders and time horizons (Mill, 1863). Contemporary consequentialist analysis in business settings typically operationalizes welfare using measures such as safety, health, economic well-being, and environmental impact, and it recommends systematic identification of benefits and harms, including indirect and downstream effects, for different groups. Within artificial intelligence governance, such analysis is frequently embedded in risk management processes, cost–benefit assessments, and monitoring and evaluation plans.

Virtue ethics shifts the focus from rule compliance or aggregate outcomes to the character and practical wisdom of agents and institutions. In Aristotle's account, moral excellence is cultivated through habituation under the guidance of practical wisdom, and a virtuous person reliably

Commented [CL1]: Part A: Theoretical Foundations
A1. Defining the three major approaches to ethics

chooses a mean between extremes as the situation requires (Aristotle, trans. Irwin, 1999/2019). Modern virtue ethics emphasizes integrity, honesty, courage, justice, and temperance, and it treats organizations as communities whose routines and incentives shape the moral character of their members (Hursthouse, 1999). In this perspective, the central question for artificial intelligence practices becomes what kind of people and what kind of organization a given technology and governance choice will form over time.

These theories differ in structure but typically converge when applied to wrongful conduct such as stealing and kidnapping. A deontological analysis condemns both acts because theft violates the duty not to take another's property and cannot be universalized without contradiction. Kidnapping violates the requirement to respect persons as ends since it overrides autonomy for instrumental purposes (Kant, 1785/2011). A utilitarian analysis condemns these acts because they predictably produce fear, injury, and social mistrust, and they undermine cooperative systems that are necessary for general welfare, even if a given instance appears to yield a local benefit (Mill, 1863). A virtue ethical analysis condemns them as displays of injustice and cowardice that damage the agent's character and erode the moral formation of the community since regular resort to coercion and deceit habituates vice rather than virtue (Aristotle, trans. Irwin, 1999/2019). These convergent judgments provide a proper calibration for later comparisons when less obvious practices, for example, manipulative design or covert data inference, are evaluated.

Commented [CL2]: A2. Baseline illustrations using clearly unethical acts

Applying the prominent three theories to artificial intelligence in business is valuable because modern governance frameworks reflect the underlying logic of these theories. European and international policy instruments highlight requirements that correspond to each perspective. The Ethics Guidelines for Trustworthy AI issued by the European Commission's High Level Expert

Commented [CL3]: A3. Why apply these theories to artificial intelligence in the workplace

Group on AI codify human agency and oversight, technical robustness, privacy and data governance, transparency, diversity and fairness, societal and environmental well-being, and accountability, which map onto duties and rights, outcome assessment, and stewardship virtues respectively (European Commission High Level Expert Group on AI, 2019). The OECD AI Principles similarly reflect democratic values and human rights while encouraging innovation, again combining duty-based and outcome-based considerations with a focus on responsible governance cultures (OECD, 2019). In the United States context, the NIST AI Risk Management Framework and NIST Special Publication 1270 on bias treat artificial intelligence risk as socio-technical, combining measurement and management of computational, human, and systemic contributors to harm, and they place consequentialist analysis within a governance context that acknowledges duties and organizational responsibility (NIST, 2023; NIST, 2022). Using deontology, consequentialism, and virtue ethics as a teaching framework, therefore, provides executives and non-specialists with a precise and complementary vocabulary that explains why contemporary standards appear as they do and how different moral questions should be weighed.

Each theory also clarifies a common workplace confusion in ways that are directly useful for managers. A deontological lens explains why covert employee monitoring or emotion inference remains unethical even when it promises efficiency gains, because such practices fail to respect autonomy and informed consent. A consequentialist lens helps teams identify indirect harms that are often overlooked in narrow business cases, including chilling effects on speech, learned helplessness from excessive automation, and second-order impacts on equity and retention. A virtue-ethical lens reminds leaders that technical controls alone are insufficient when day-to-day habits reward opacity and speed over honesty and prudence, as culture and character shape how any tool is actually used. Taken together, these perspectives prepare the analysis that follows,

with deontology guiding questions of notice, consent, and appeal, consequentialism structuring lifecycle cost–benefit and externality assessment, and virtue ethics orienting leadership transparency and practical wisdom across cost–benefit and risk management, hiring, performance reviews, stakeholder communication, and marketing.

A consequentialist approach to artificial intelligence in the enterprise requires leaders to evaluate not only direct value, such as productivity gains and cycle time reductions, but also externalities, such as disparate impact, privacy harms, employee well-being, environmental costs associated with compute, misinformation spillover, and brittle dependencies in supply chains or models. The NIST AI Risk Management Framework provides a structure for governance, mapping, measurement, and management across the development and deployment lifecycle, and NIST SP 1270 differentiates sources of bias as computational, human, and systemic, encouraging mitigation that goes beyond technical adjustments to include organizational change (NIST, 2023; NIST, 2022). The ISO/IEC 42001 standard specifies requirements for a certifiable AI management system, integrating policy, risk registers, impact assessments, supplier oversight, and human-in-the-loop design into a coherent governance backbone that can be audited and continuously improved (ISO/IEC, 2023). A deontological screen complements these analyses by identifying practices that are off limits regardless of net benefit. For example, deployments that covertly infer sensitive attributes or that apply emotion recognition to employees without informed consent undermine duties of respect and honesty, and many such practices are restricted or prohibited in the EU AI Act's risk framework (European Union, 2024). A virtue-ethical orientation then asks whether the leadership team cultivates practical wisdom by documenting reasons, acknowledging uncertainty, inviting critique, and modeling integrity in communication. Research on leadership transparency in AI-intensive organizations suggests that

Commented [CL4]: Part B: Application to CEO Success
B1. Cost–benefit and risk assessment

when leaders fail to explain the rationale for technology choices, trust erodes and ethical agency narrows, which undermines the quality of decision-making (Edmondson, 2019; Sejera & Bocarnea, 2022).

In practical terms, executives should translate these principles into explicit decision thresholds that determine whether a system can advance to pilot and production. Before launch, the chief executive should require a demonstration of subgroup performance within agreed-upon tolerances, for example, by showing that differences in false positive and false negative rates across protected groups under an equalized odds analysis remain within a narrow band. The probability estimates are well-calibrated for each group, with calibration slopes close to one and minimal intercept bias. Environmental due diligence should be incorporated into the risk register by estimating training and inference energy use in kilowatt-hours, converting those estimates to carbon dioxide equivalents using region-specific grid carbon intensity, and recording the effects of data center efficiency and model size choices on the footprint. Technical strategies should be selected to meet these thresholds with the least necessary resource use, including model distillation to smaller architectures, quantization where accuracy permits, and retrieval-augmented generation that reduces reliance on huge models by grounding outputs in curated knowledge sources.

Core competencies for the chief executive include risk literacy with reference to NIST and ISO frameworks, an auditability mindset, an understanding of fairness tradeoffs, and the institutional humility to pause or halt deployments when duties or documented risks require it. These competencies should be visible in practice. The chief executive should chair or regularly attend a cross-functional AI governance forum that reviews subgroup performance tables, calibration plots, and error analyses, and the chief executive should insist that vendor contracts include audit

rights, documentation deliverables such as Model Cards and Datasheets, and clear routes for remediation and rollback if material risks emerge.

Empirical research demonstrates that algorithms can exacerbate inequities when proxy variables reflect underlying structural disparities. A widely used health care tool assigned lower risk scores to Black patients than to White patients with similar health needs because the algorithm used cost as a proxy for need, encoding historical differences in access to care (Obermeyer, Powers, Vogeli, & Mullainathan, 2019). Reviews of algorithmic hiring practices reveal opaque vendor claims, inconsistent validation, and unresolved fairness trade-offs, recommending rigorous documentation and audits over reliance on vendor assurances (Raghavan, Barocas, Kleinberg, & Levy, 2020). Regulatory bodies have also emphasized the risk of discrimination in employment screening, underscoring the importance of thorough documentation and human review. A deontological analysis of company-wide hiring, therefore, requires informed notice, meaningful appeal, and non-discrimination as matters of duty, which in turn implies controls against covert psychometric inference and non-consensual affective analysis. A consequentialist analysis weighs the efficiency benefits against systemic externalities, including false negatives that systematically exclude protected groups, which can lead to legal exposure, reputational damage, and social mistrust. Documentation practices should make scope, limitations, and subgroup performance visible, enabling both internal and external oversight (Gebru et al., 2021; Mitchell et al., 2019). A virtue-ethical perspective reframes hiring as a formation practice that shapes the organization's character, since heavy reliance on opaque scoring can estrange managers from the virtues of fair deliberation and empathy. Research in information systems suggests that sustained performance improvements from artificial intelligence require complementary tangible, human, and organizational resources. Without these resources, organizations risk superficial adoption

Commented [CL5]: B2. Hiring processes at the company-wide level

that erodes capability and trust rather than augmenting it (Mikalef & Gupta, 2021). Work design research also indicates that job quality depends on dimensions such as autonomy, mastery, stimulation, relational support, and tolerable demands, and that automation, which degrades these dimensions, will harm engagement and equity in the long term (Parker & Knight, 2024). For these reasons, the recommended role of artificial intelligence in hiring is one of support rather than determination. Artificial intelligence is well-suited to sourcing, deduplication, and flagging qualified candidates, as well as screening with validated instruments whose subgroup performance is disclosed and monitored. At the same time, evaluative uses that infer fit from facial or vocal cues or opaque instructions should be avoided. Senior leaders need the ability to interpret audit reports, to insist on documentation artifacts in procurement, and to align vendor oversight with ISO/IEC 42001 controls.

There is also a stronger-case argument for adoption within each moral tradition when specific guardrails are in place. From a deontological perspective, consistent and criteria-based screening can advance the duties of fairness when candidates receive clear notice about automated processing and a meaningful avenue for appeal to a human decision-maker. From a consequentialist perspective, validated tools can reduce the time to hire, improve matching between applicants and roles, and widen talent pools when screening correlates with job-relevant outcomes rather than proxies, provided that monitoring detects and corrects disparate error rates. From a virtue-ethical perspective, decision support systems can scaffold more reflective deliberation by surfacing relevant evidence, reducing reliance on irrelevant heuristics, and prompting managers to record reasons for decisions in ways that model practical wisdom for the organization. To institutionalize these benefits, the firm should maintain a model registry for all hiring algorithms, convene a cross-functional review committee that includes HR, legal, data

science, and diversity officers, require vendor documentation at procurement and at every significant model update, and provide independent appeal channels for candidates that are separate from the hiring team that used the model.

Artificial intelligence that summarizes goals, surfaces work examples, or assists with pattern detection can improve feedback quality and timeliness, providing potential welfare gains in consequentialist terms. These potential gains are offset by risks of misclassification, especially in hybrid and remote roles, and by penalizing caregiving or atypical work patterns that do not correlate with actual value creation. A deontological approach to performance management foregrounds employees' rights to know and contest how they are evaluated and requires that any monitoring be proportionate and transparent. A virtue perspective emphasizes development rather than scoring and treats performance evaluation as an opportunity to cultivate excellence through coaching and narrative judgment. Experimental research suggests that employees are less likely to comply with unethical instructions from artificial systems than with those from humans, implying that organizations cannot simply transfer moral authority to automated systems; justification and explanation remain essential components of legitimate evaluation (Schlager, Bhattacharjee, & Häubl, 2023). Consequently, the appropriate role for artificial intelligence in performance management is to support human judgment rather than replace it. Case-level explainability, visible criteria, and manager sign-off remain critical for ensuring fairness and trust.

Commented [CL6]: B3. Annual reports and performance reviews

Case-level explainability should be specific enough for both managers and employees to understand and contest a proposed rating. A defensible system will demonstrate the relative contribution of objective metrics, such as goal attainment and timeliness, as well as qualitative evidence, including peer and client feedback, and modeled inferences drawn from patterns in the

work record. Both the manager and the employee should be able to view the underlying evidence that supports the recommendation, annotate it, request reconsideration, and have those interactions logged so that the organization can audit how the model is used in practice.

AI-augmented productivity scoring and monitoring require special scrutiny. Keyboard or mouse telemetry, application focus logs, and webcam analysis may appear to promise efficiency; yet, a deontological analysis often finds such surveillance disproportionate to the purpose and disrespectful of autonomy and privacy, especially when collection continues outside core work tasks or without explicit consent. A consequentialist analysis must account for morale costs, chilling effects on initiative, higher attrition rates, and potential labor relations and legal risks, which often outweigh short-term output changes. Where monitoring is necessary for safety or compliance, it should be narrowly tailored, time-bound, and transparent, with outcome-oriented measures, such as completed work against agreed-upon goals, preferred over continuous monitoring of behavior. Firms should minimize collection, aggregate, and anonymize telemetry where feasible, and provide employees with visibility into what is captured and how it is used.

Generative artificial intelligence can enhance clarity and access by assisting with drafting, summarizing, and translation, which can increase welfare for diverse stakeholder groups. The same capability can be misused by simulating stakeholder voices, creating synthetic consensus, or producing persuasive content that obscures its origin, thereby violating the duties of truth-telling and respect for persons. A consequentialist view acknowledges that deceptive communications erode trust and can lead to legal and reputational crises. A virtue ethics lens highlights the moral quality of transparent communication in organizations that heavily rely on artificial intelligence, emphasizing that leaders who clearly explain uncertainty and reasons promote psychological safety, and that such habits foster integrity over time (Edmondson, 2019;

Commented [CL7]: B4. Communication with stakeholders

Floridi & Cowls, 2019). Governance instruments, including IEEE's Ethically Aligned Design and the United Nations' guidance on ethical artificial intelligence, emphasize transparency, accountability, and human rights in ways that can be incorporated into corporate communication policies (IEEE Global Initiative, 2019; United Nations, 2021). These considerations favor assistive uses that increase accessibility and accuracy while prohibiting deception and manipulation.

To make these norms operational, firms should adopt specific governance practices for regulated and investor communications. AI-assisted disclosures should be internally tagged, reviewed by legal and finance for factual accuracy and compliance, and managed through version control so that decision provenance can be reconstructed. Public-facing materials should adhere to a content-provenance policy that prohibits the use of synthetic personas in investor or policy engagement, and where appropriate, applies watermarking or cryptographic signatures to indicate the origin of the content. These measures help ensure that assistive use improves clarity without crossing into deception.

Artificial intelligence can improve relevance, accessibility, and product discovery for customers when marketing content and experiences are personalized responsibly. It can also intensify manipulative design, especially when agentic chatbots or conversational agents leverage user data to exploit cognitive vulnerabilities. A deontological analysis treats manipulation that undermines autonomy as disrespectful regardless of its profitability. A consequentialist analysis warns of downstream harms to financial well-being and health from aggressive targeting and nudging. A virtue-ethical analysis asks whether the organization is becoming more honest and respectful in its engagement with customers, or whether it is cultivating a culture that normalizes exploitation. Research on dark patterns in human–robot interaction and persuasive interfaces

Commented [CL8]: B5. Marketing products and services

highlights the need for an explicit review of designs that may exploit users' cognitive constraints for commercial gain. Policy signals in the EU AI Act, including restrictions on specific manipulative systems and biometric categorization, further justify conservative approaches to affective and inferential marketing practices (European Union, 2024). The practical conclusion is that artificial intelligence should be used to enhance clarity, accessibility, and matching, while manipulative patterns should be identified and opt-out mechanisms made transparent.

Beyond conversational agents, risks also arise in dynamic pricing and hyper-segmented offers. Responsible personalization should avoid making sensitive inferences about health, sexuality, religion, or political beliefs unless explicit and informed consent is obtained; even then, such uses should undergo heightened review. Children's data and health-related marketing deserve special treatment, with stricter limits on profiling, a bias toward contextual rather than behavioral targeting, and precise parental controls. Compatibility rules for personalization should include simple, persistent opt-out mechanisms across channels, frequency caps that limit repeated persuasion attempts, and mandatory human review for campaigns targeting potentially vulnerable populations, such as minors, the elderly, or those experiencing recent financial distress.

Across finance and fintech, recent research converges on the view that algorithmic performance is inseparable from context, incentives, and institutional design. Studies of credit markets show that advanced models can sharpen risk assessment while also re-inscribing or widening disparities if firms optimize on signals that correlate with protected characteristics or with historically unequal access to opportunity. In mortgage lending, machine-learning pricing and underwriting have been shown to produce predictably unequal outcomes across borrower groups even when explicit protected attributes are excluded from the feature set, a pattern consistent

Commented [CL9]: Part C: Evidence and Research Integration
C1. Themes and patterns across literature

with lenders' use of highly predictive but distributionally skewed proxies (Fuster, Goldsmith-Pinkham, Ramadorai, & Walther, 2022). In consumer credit more broadly, evidence from the most significant U.S. market suggests that fintech lenders have not eliminated discrimination relative to traditional institutions; price differences persist for similarly situated applicants, indicating that algorithmic improvements in screening do not automatically produce equitable pricing (Bartlett, Morse, Stanton, & Wallace, 2022). At the same time, digital footprint variables used by online lenders can match or exceed the information content of bureau scores and expand access when used responsibly, underscoring the double-edged nature of alternative data in underwriting (Berg, Burg, Gombović, & Puri, 2020). These findings suggest that bias is a socio-technical phenomenon rooted not only in data and code, but also in business objectives, market power, and governance choices, rather than a purely statistical artifact.

Accountability and explainability have therefore become central to the credible use of AI in credit, payments, and wealth management. Financial institutions increasingly pair model documentation with monitoring and audit routines; however, the literature emphasizes that artifacts are effective only when linked to decision rights and quality-assurance workflows. A multi-vendor empirical program by FinRegLab finds that explainability tools can support supervisory obligations and model risk functions, yet their usefulness depends on alignment between the explanation technique and the specific decision context (e.g., underwriting vs. portfolio management) and on disciplined operational use over time (FinRegLab, 2023). Complementary surveys in 2024–2025 map the rapidly expanding catalog of financial XAI methods and argue that credit-risk and underwriting have matured most, while use cases like portfolio optimization, internet finance, and fraud detection remain comparatively under-explored from an explainability standpoint (Černevičienė et al., 2024; Yeo, Lee, & Kim,

2025). Taken together, this body of work frames explainability not as a mere transparency add-on but as a control function that conditions how models are trained, tested, and governed across the lifecycle.

Regulatory developments over the last five years reinforce this shift from "black-box tolerance" to demonstrable accountability. In the European Union, the Artificial Intelligence Act classifies credit scoring and creditworthiness assessment as "high-risk," triggering obligations for risk management, data and model documentation, and post-market monitoring (European Union, 2024). In the United States, the Consumer Financial Protection Bureau clarified that creditors relying on complex models must still provide specific, accurate reasons for adverse actions under ECOA and Regulation B, regardless of the technology used, thereby rejecting generic or non-specific explanations (CFPB, 2022; CFPB, 2024). Bank supervisors have simultaneously signaled that existing model-risk frameworks apply to AI, updating model-risk handbooks and issuing joint third-party risk guidance that squarely encompasses model-as-a-service and other SaaS vendor arrangements common in fintech ecosystems (OCC, 2021; Federal Banking Agencies, 2023). Internationally, prudential and standard-setting bodies have highlighted explainability as a supervisory priority in financial services, linking it to validation, challenge, and consumer-protection outcomes (BIS-FSI, 2024).

A recurring theme in this literature is the fragility of fairness and performance under even small shifts in data, labeling, and deployment context. Institutions report that subgroup error rates can drift as macroeconomic conditions, product mixes, or customer acquisition channels evolve, and that one-time pre-launch validations are poor substitutes for continuous monitoring linked to escalation and rollback. Research syntheses in finance emphasize the distribution shift and the need to pair statistical parity checks with business-process audits to ensure that metric-level

improvements translate into better outcomes for affected groups (Das, Stanton, & Sundaram, 2023; FinRegLab, 2023). In practice, this means treating model updates as high-risk changes, subject to change control, testing explanations for stability, and stress-testing models against representative slices of the portfolio that reflect geographic, demographic, and channel heterogeneity.

Beyond credit, two adjacent domains illustrate both promise and caution. In wealth management, evidence from a large-scale robo-advising rollout shows that automated advice can improve diversification and adherence to goals relative to baseline investor behavior; however, the benefits depend on careful product design, disclosure, and the avoidance of one-size-fits-all templating (D'Acunto, Prabhala, & Rossi, 2019). In capital markets, large language models can extract tradable signals from news, producing statistically significant predictability in initial and short-horizon reactions. However, these effects raise questions about robustness, market impact, and the governance of model-driven trading strategies (Lopez-Lira & Tang, 2023). Meanwhile, payments and compliance functions have turned to machine learning for fraud and AML monitoring. Early peer-reviewed evidence indicates efficiency gains and a reduction in false positives when models are paired with strong oversight and investigator workflows, rather than being operated as fully autonomous systems (JMLC, 2025). Together, these strands demonstrate that the adoption of fintech and SaaS AI is most successful when technical advances are integrated into institutional arrangements that ensure explainability, contestability, and continuous control.

Finally, the platformed nature of financial AI, where many banks consume risk models as SaaS or via third-party APIs, means that vendor governance has become a core competency. Interagency third-party guidance in the United States now expects banks to extend model risk

disciplines across contractual boundaries, including rights to documentation, testing artifacts, performance dashboards, and remediation. In contrast, the EU regime places direct obligations on providers of high-risk systems. This evolution aligns the literature's central claim with practice: the quality of financial AI is determined as much by contractual, organizational, and regulatory design as by algorithms themselves.

The research base benefits from multi-method evaluation, including field audits of deployed systems, lifecycle governance frameworks, empirical documentation practices, and policy analyses. This breadth increases external validity and policy relevance; however, the pace of change in both technology and law creates a risk that findings or guidance will become outdated, necessitating continuous monitoring and revision. Many fairness studies rely on specific proxies or outcomes, and their portability across domains is uncertain, suggesting the need for replication and domain-specific validation. Documentation practices can devolve into checklists unless they are linked to governance incentives, such as certification or regulatory obligations. Behavioral studies on human responses to algorithmic authority often rely on online samples or artificial tasks, which limit generalizability, although directional insights remain informative.

Commented [CL10]: C2. Methodological strengths and limitations

Executives should be alert to several threats to validity that recur across applied studies. Improvements in metric-level fairness do not always translate into real-world gains for affected groups when organizational processes downstream of the model remain unchanged. Data sets can harbor spurious correlations that drive apparent accuracy without reflecting causal relationships, making models brittle under modest shifts in context. Distribution shifts in production, whether due to seasonality, policy changes, or user adaptation to the system, can quickly erode performance and subgroup parity unless there is active monitoring, drift detection, and a defined process for rollback or recalibration.

A further limitation concerns documentation drift and audit constraints. Model Cards and Datasheets can become stale when teams update data, features, or parameters without synchronized changes to artifacts, which undermines their value as living records. Effective programs, therefore, connect documentation updates to change-control triggers and require signoff before promotion to production. Third-party auditors can face information-access limits that weaken rigor, including incomplete access to training data, proprietary feature engineering pipelines, or redacted performance dashboards. Firms should anticipate these constraints by negotiating audit-friendly vendor contracts, preparing secure data-room access for auditors, and pre-defining safe-harbor procedures that balance confidentiality with the need for meaningful verification.

Several gaps exist for future research. First, virtue ethics requires further operationalization in organizational settings, including the development of measures for practical wisdom and virtue cultivation in the governance of artificial intelligence. Second, more attention is needed to adoption by small and medium enterprises, which often lack the resources to implement heavy documentation and audit regimes and would benefit from lightweight and scalable methods. Third, scholarship remains concentrated in Western contexts and requires expansion to include Global South perspectives and culturally contextual notions of fairness and rights, as well as the environmental impacts associated with the production and end-of-life storage of e-waste. Fourth, longitudinal research is needed on the job-quality impacts of artificial intelligence, utilizing frameworks such as the SMART model to connect design decisions to autonomy, mastery, strain, and retention.

To address these needs, future work could employ randomized or staggered rollouts to assess the causal impact of mitigation techniques, and it could build panel datasets that track employee

Commented [CL11]: C3. Gaps and directions for future inquiry

outcomes over time, linking work design changes to retention, performance, and well-being. Ethical decision-making can be integrated into a practical sequencing of executive judgment. A deontic screen should be applied first to identify practices that are impermissible because they violate rights, duties, or applicable law, such as non-consensual inferences or manipulative affective systems, which may be restricted by instruments like the EU AI Act (European Union, 2024). A consequentialist analysis should then evaluate full-lifecycle benefits and harms, including externalities, using risk management frameworks such as those provided by NIST to structure hazard identification, measurement, monitoring, and response (NIST, 2023). A virtue-ethical reflection should consider whether the choice will form the organization into one that is more transparent, fair, and trustworthy, and whether leadership practices such as decision logs, admission of uncertainty, and invitation of red-team critique are in place to cultivate practical wisdom over time. Documentation and audits should be treated as go-live gates rather than as afterthoughts, and meaningful human appeal should be established for high-stakes outcomes in line with international principles of trustworthy artificial intelligence (Mitchell et al., 2019; Gebru et al., 2021; Raji et al., 2020; European Commission High-Level Expert Group on AI, 2019). In concrete terms, advancement to deployment should be conditional on the presence of specific artifacts: complete Datasheets for the datasets employed, Model Cards that disclose the model's scope and subgroup performance, end-to-end audit reports that trace accountability across the lifecycle, and signed human-oversight plans that specify escalation and recourse. To implement this sequence, senior leadership must demonstrate fluency with diagnostic tools, such as confusion matrices and calibration plots, that reveal performance and parity. They must also be able to set vendor audit clauses and manage incident response, including timely user notification and remediation when material risks or harms are identified.

Commented [CL12]: Integrating the Big Three into a CEO decision approach

Ultimately, more education and time will need to be set aside to understand the tools being implemented, compared to prior intelligent software systems.

Each theory can, under the right conditions, support the deployment of artificial intelligence. Deontology recognizes the duties of beneficence and non-maleficence. It therefore favors systems that prevent harm and protect individuals, such as fraud detection or safety monitoring, provided that privacy, consent, and explanation duties are respected (Kant, 1785/2011). Consequentialism supports the use of artificial intelligence when there is a clear and monitored net welfare improvement, highlighting the importance of lifecycle governance, documentation, and post-deployment monitoring to confirm expected benefits and identify unanticipated harms (NIST, 2023; ISO/IEC, 2023). Virtue ethics supports artificial intelligence that cultivates organizational virtues, such as fairness-by-design, transparency rituals, and prudent stop rules, and it cautions against deployments that corrupt character by normalizing manipulation or deskilling judgment (Aristotle, trans. Irwin, 1999/2019). Put more actionably, deontology's green lights include consent-respecting, rights-preserving assistive systems, while its red lines include deception, non-consensual surveillance, and discriminatory profiling; consequentialism's green lights include deployments with verified utility gains and robust monitoring, while its red lines include systems whose externalities remain unmitigated or shift harms onto vulnerable groups; virtue ethics' green lights include tools that reinforce honesty, fairness, and prudence in everyday practice, while its red lines include tools that habituate opacity, manipulation, or moral abdication.

Commented [CL13]: Counter-points and synthesis by theory

Artificial intelligence is most suitable for back-office acceleration, synthesis of complex documents, coding assistance with guardrails, and customer support triage, where human review supervises outputs and limitations are disclosed. It should be used cautiously, or not at all, for

Commented [CL14]: CEO takeaway on where and how much artificial intelligence to use

emotion inference, covert personalization that exploits vulnerabilities, surveillance-intensive productivity scoring, and high-stakes automated decisions without an appeal path. Building organizational capability involves governance literacy with ISO/IEC 42001 and EU requirements, risk and audit literacy with NIST frameworks, stewardship of data through Datasheets and model documentation through Model Cards, attention to work design using frameworks such as SMART, and cultivation of leadership habits that make reasons and uncertainties visible to the workforce and to stakeholders (ISO/IEC, 2023; NIST, 2023; Gebru et al., 2021; Mitchell et al., 2019; Parker & Knight, 2024). To institutionalize these practices, the organization should implement executive training on AI ethics, establish a recurring governance cadence through a standing cross-functional review, and incorporate concrete AI ethics outcomes into executive scorecards, ensuring accountability is visible and durable.

Deontology, consequentialism, and virtue ethics are not merely historical doctrines but operational tools for modern artificial intelligence governance in business. Deontology protects dignity and rights through non-negotiable duties, consequentialism aligns deployment with welfare through structured risk management and monitoring, and virtue ethics ensures that technology practice forms people and institutions in the direction of integrity. Chief executives who internalize all three perspectives and institutionalize them through standards, audits, documentation, and leadership habits are positioned to capture the value of artificial intelligence while earning and maintaining stakeholder trust.

References

Aristotle. (1999/2019). *Nicomachean ethics* (T. Irwin, Trans.). Hackett.

Bank for International Settlements—Financial Stability Institute. (2024). *Managing explanations: How regulators can address AI explainability* (FSI Insights on policy implementation No. 24). https://www.bis.org/fsi/fsipapers24.htm

Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2022). Consumer-lending discrimination in the FinTech era. *Journal of Financial Economics, 143*(1), 30–56. https://doi.org/10.1016/j.jfineco.2021.08.011

Berg, T., Burg, V., Gombović, A., & Puri, M. (2020). On the rise of FinTechs: Credit scoring using digital footprints. *Review of Financial Studies, 33*(7), 2845–2897. https://doi.org/10.1093/rfs/hhz099

Černevičienė, J., Grigonis, V., & Valatka, M. (2024). Explainable artificial intelligence in finance: A systematic review. *Artificial Intelligence Review*. https://doi.org/10.1007/s10462-024-10854-8

Consumer Financial Protection Bureau. (2022). *Circular 2022-03: Adverse action notification requirements in connection with credit decisions based on complex algorithms*. https://www.consumerfinance.gov/compliance/circulars/circular-2022-03-adverse-action-notification-requirements-in-connection-with-credit-decisions-based-on-complex-algorithms/

Consumer Financial Protection Bureau. (2024). *Consumer Financial Protection Circular 2023-03: Adverse action notification requirements and proper use of sample forms*. *Federal Register,*

89(75). https://www.federalregister.gov/documents/2024/04/17/2024-08003/consumer-financial-protection-circular-2023-03-adverse-action-notification-requirements-and-proper

D'Acunto, F., Prabhala, N., & Rossi, A. G. (2019). The promises and pitfalls of robo-advising. *Review of Financial Studies, 32*(5), 1983–2020. https://doi.org/10.1093/rfs/hhy100

Das, S., Stanton, R., & Sundaram, R. (2023). Algorithmic fairness. *Annual Review of Financial Economics, 15*, 455–480. https://doi.org/10.1146/annurev-financial-110921-125930

Edmondson, A. C. (2019). *The fearless organization: Creating psychological safety in the workplace for learning, innovation, and growth*. Wiley.

European Commission High-Level Expert Group on AI. (2019). *Ethics guidelines for trustworthy AI*.

European Union. (2024). *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. Summary: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai ; Annex III: https://artificialintelligenceact.eu/annex/3/

Federal Banking Agencies (Board of Governors of the Federal Reserve System; FDIC; OCC). (2023). *Interagency guidance on third-party relationships: Risk management*. https://www.occ.gov/news-issuances/news-releases/2023/nr-ia-2023-53a.pdf

FinRegLab. (2023). *Explainability and fairness in machine learning for credit underwriting: Policy analysis and industry survey*. https://finreglab.org/research/explainability-fairness-credit-underwriting/

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

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2022). Predictably unequal? The effects of machine learning on credit markets. *Journal of Finance, 77*(4), 2049–2097. https://doi.org/10.1111/jofi.13090

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM, 64*(12), 86–92.

Hursthouse, R. (1999). *On virtue ethics*. Oxford University Press.

IEEE Global Initiative. (2019). *Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems* (Version 2). IEEE Standards Association.

ISO/IEC. (2023). *ISO/IEC 42001: Artificial intelligence management system*. International Organization for Standardization.

Journal of Money Laundering Control. (2025). Enhancing AML compliance: A machine-learning approach to transaction monitoring. *Journal of Money Laundering Control, 28*(3). https://www.emerald.com/insight/publication/issn/1368-5201

Kant, I. (1785/2011). *Groundwork of the metaphysics of morals* (M. Gregor & J. Timmermann, Eds.). Cambridge University Press.

Lopez-Lira, A., & Tang, Y. (2023). Can ChatGPT forecast stock price movements? Return predictability and large language models. *arXiv preprint* arXiv:2304.07619. https://arxiv.org/abs/2304.07619

Mathur, A., Acar, G., Friedman, M. G., Lucherini, E., Mayer, J., Chetty, M., & Narayanan, A. (2019). Dark patterns at scale: Findings from a crawl of 11K shopping websites. *Proceedings of the ACM on Human-Computer Interaction, 3*(CSCW), Article 81, 1–32. https://doi.org/10.1145/3359183

Mill, J. S. (1863). *Utilitarianism*. (Various editions).

Mikalef, P., & Gupta, M. (2021). Artificial intelligence capability: Conceptualization, measurement calibration, and empirical study on its impact on organizational creativity and firm performance. *Information & Management, 58*, 103434.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., … Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220–229). ACM.

National Institute of Standards and Technology. (2022). *Towards a standard for identifying and managing bias in AI* (NIST Special Publication 1270).

National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. https://www.nist.gov/itl/ai-risk-management-framework

OECD. (2019). *Recommendation of the Council on Artificial Intelligence*.

Office of the Comptroller of the Currency. (2021/2025 update). *Comptroller's Handbook: Model risk management* (Version 1.0; updated March 2025 notes). https://www.occ.treas.gov/publications-and-resources/publications/comptrollers-handbook/files/model-risk-management/pub-ch-model-risk.pdf

Parker, S. K., & Knight, C. (2024). The SMART model of work design: A higher order structure to help see the wood from the trees. *Human Resource Management, 63*(2), 265–291.

Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 469–481). ACM.

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

Ross, W. D. (1930/2002). *The right and the good* (P. Stratton-Lake, Ed.). Oxford University Press.

Schlager, T., Bhattacharjee, A., & Häubl, G. (2023). Employees adhere more to unethical instructions from humans than AI. *Journal of Experimental Psychology: Applied*. Advance online publication.

United Nations. (2021). *A framework for ethical AI at the United Nations*.

Yeo, W.-J., Lee, J., & Kim, M. (2025). A comprehensive review on financial explainable AI. *Artificial Intelligence Review*. https://doi.org/10.1007/s10462-024-11077-7

Note: Where publisher landing pages or official summaries are cited for legislation and handbooks (e.g., EU AI Act; OCC Handbook), the citations point to authoritative portals that host the operative texts or official overviews.

This paper utilized AI-augmented decision-making, specifically Grammarly (2025, https://www.grammarly.com/grammar), to refine grammar and tone. Additionally, ChatGPT (https://openai.com/chatgpt/overview/) was used to help compare the final draft of the paper to the prompt for coherence and completeness of the answer.

The final submission reflects the author's critical assessment and revision of AI-generated suggestions. An ethical framework was employed to assess the benefits and risks associated with this action. All original writing is the Author’s, unless cited otherwise.