

THE CAPCO INSTITUTE
JOURNAL
OF FINANCIAL TRANSFORMATION

WORLD
OF
FINANCE
AND
TECHNOLOGY

a **wipro** company



GenAI

2024/2025 EDITION

THE CAPCO INSTITUTE

JOURNAL OF FINANCIAL TRANSFORMATION

RECIPIENT OF THE APEX AWARD FOR PUBLICATION EXCELLENCE

Editor

Shahin Shojai, Global Head, Capco Institute

Advisory Board

Lance Levy, Strategic Advisor

Owen Jelf, Partner, Capco

Suzanne Muir, Partner, Capco

David Oxenstierna, Partner, Capco

Editorial Board

Franklin Allen, Professor of Finance and Economics and Executive Director of the Brevan Howard Centre, Imperial College London and Professor Emeritus of Finance and Economics, the Wharton School, University of Pennsylvania

Philippe d'Arvisenet, Advisor and former Group Chief Economist, BNP Paribas

Rudi Bogni, former Chief Executive Officer, UBS Private Banking

Bruno Bonati, Former Chairman of the Non-Executive Board, Zuger Kantonalbank, and President, Landis & Gyr Foundation

Dan Breznitz, Munk Chair of Innovation Studies, University of Toronto

Urs Birchler, Professor Emeritus of Banking, University of Zurich

Elena Carletti, Professor of Finance and Dean for Research, Bocconi University, Non-Executive Director, UniCredit S.p.A.

Lara Cathcart, Associate Professor of Finance, Imperial College Business School

Géry Daeninck, former CEO, Robeco

Jean Dermine, Professor of Banking and Finance, INSEAD

Douglas W. Diamond, Merton H. Miller Distinguished Service Professor of Finance, University of Chicago

Elroy Dimson, Emeritus Professor of Finance, London Business School

Nicholas Economides, Professor of Economics, New York University

Michael Enthoven, Chairman, NL Financial Investments

José Luis Escrivá, President, The Independent Authority for Fiscal Responsibility (AIReF), Spain

George Feiger, Pro-Vice-Chancellor and Executive Dean, Aston Business School

Gregorio de Felice, Head of Research and Chief Economist, Intesa Sanpaolo

Maribel Fernandez, Professor of Computer Science, King's College London

Allen Ferrell, Greenfield Professor of Securities Law, Harvard Law School

Peter Gomber, Full Professor, Chair of e-Finance, Goethe University Frankfurt

Wilfried Hauck, Managing Director, Statera Financial Management GmbH

Pierre Hillion, The de Picciotto Professor of Alternative Investments, INSEAD

Andrei A. Kirilenko, Reader in Finance, Cambridge Judge Business School, University of Cambridge

Katja Langenbacher, Professor of Banking and Corporate Law, House of Finance, Goethe University Frankfurt

Mitchel Lenson, Former Group Chief Information Officer, Deutsche Bank

David T. Llewellyn, Professor Emeritus of Money and Banking, Loughborough University

Eva Lomnicka, Professor of Law, Dickson Poon School of Law, King's College London

Donald A. Marchand, Professor Emeritus of Strategy and Information Management, IMD

Colin Mayer, Peter Moores Professor of Management Studies, Oxford University

Francesca Medda, Professor of Applied Economics and Finance, and Director of UCL Institute of Finance & Technology, University College London

Pierpaolo Montana, Group Chief Risk Officer, Mediobanca

John Taysom, Visiting Professor of Computer Science, UCL

D. Sykes Wilford, W. Frank Hipp Distinguished Chair in Business, The Citadel

CONTENTS

TECHNOLOGY

08 Mindful use of AI: A practical approach

Magnus Westerlund, Principal Lecturer in Information Technology and Director of the Laboratory for Trustworthy AI, Arcada University of Applied Sciences, Helsinki, Finland

Elisabeth Hildt, Affiliated Professor, Arcada University of Applied Sciences, Helsinki, Finland, and Professor of Philosophy and Director of the Center for the Study of Ethics in the Professions, Illinois Institute of Technology, Chicago, USA

Apostolos C. Tsolakis, Senior Project Manager, Q-PLAN International Advisors PC, Thessaloniki, Greece

Roberto V. Zicari, Affiliated Professor, Arcada University of Applied Sciences, Helsinki, Finland

14 Understanding the implications of advanced AI on financial markets

Michael P. Wellman, Lynn A. Conway Collegiate Professor of Computer Science and Engineering University of Michigan, Ann Arbor

20 Auditing GenAI systems: Ensuring responsible deployment

David S. Krause, Emeritus Associate Professor of Finance, Marquette University

Eric P. Krause, PhD Candidate – Accounting, Bentley University

28 Innovating with intelligence: Open-source Large Language Models for secure system transformation

Gerhardt Scriven, Executive Director, Capco

Tony Moenicke, Senior Consultant, Capco

Sebastian Ehrig, Senior Consultant, Capco

38 Multimodal artificial intelligence: Creating strategic value from data diversity

Cristián Bravo, Professor, Canada Research Chair in Banking and Insurance Analytics, Department of Statistical and Actuarial Sciences, Western University

46 GenAI and robotics: Reshaping the future of work and leadership

Natalie A. Pierce, Partner and Chair of the Employment and Labor Group, Gunderson Dettmer

ORGANIZATION

56 How corporate boards must approach AI governance

Arun Sundararajan, Harold Price Professor of Entrepreneurship and Director of the Fubon Center for Technology, Business, and Innovation, Stern School of Business, New York University

66 Transforming organizations through AI: Emerging strategies for navigating the future of business

Feng Li, Associate Dean for Research and Innovation and Chair of Information Management, Bayes Business School (formerly Cass), City St George's, University of London

Harvey Lewis, Partner, Ernst & Young (EY), London

74 The challenges of AI and GenAI use in the public sector

Albert Sanchez-Graells, Professor of Economic Law, University of Bristol Law School

78 AI safety and the value preservation imperative

Sean Lyons, Author of Corporate Defense and the Value Preservation Imperative: Bulletproof Your Corporate Defense Program

92 Generative AI technology blueprint: Architecting the future of AI-infused solutions

Charlotte Byrne, Managing Principal, Capco

Thomas Hill, Principal Consultant, Capco

96 Unlocking AI's potential through metacognition in decision making

Sean McMinn, Director of Center for Educational Innovation, Hong Kong University of Science and Technology

Joon Nak Choi, Advisor to the MSc in Business Analytics and Adjunct Associate Professor, Hong Kong University of Science and Technology

REGULATION

104 Mapping GenAI regulation in finance and bridging the gaps

Nydia Remolina, Assistant Professor of Law, and Fintech Track Lead, SMU Centre for AI and Data Governance, Singapore Management University

112 Board decision making in the age of AI: Ownership and trust

Katja Langenbucher, Professor of Civil Law, Commercial Law, and Banking Law, Goethe University Frankfurt

122 The transformative power of AI in the legal sector: Balancing innovation, strategy, and human skills

Eugenia Navarro, Lecturer and Director of the Legal Operations and Legal Tech Course, ESADE

129 Remuneration on the management board in financial institutions: Current developments in the framework of supervisory law, labor law, behavioral economics and practice

Julia Redenius-Hövermann, Professor of Civil Law and Corporate Law and Director of the Corporate Governance Institute (CGI) and the Frankfurt Competence Centre for German and Global Regulation (FCCR), Frankfurt School of Finance and Management

Lars Hinrichs, Partner at Deloitte Legal Rechtsanwaltsgesellschaft mbH (Deloitte Legal) and Lecturer, Frankfurt School of Finance and Management



CAPCO CEO WELCOME

DEAR READER,

Welcome to our very special 60th edition of the Capco Journal of Financial Transformation.

The release of this milestone edition, focused on GenAI, reinforces Capco's enduring role in leading conversations at the cutting edge of innovation, and driving the trends shaping the financial services sector.

There is no doubt that GenAI is revolutionizing industries and rapidly accelerating innovation, with the potential to fundamentally reshape how we identify and capitalize on opportunities for transformation.

At Capco, we are embracing an AI infused future today, leveraging the power of GenAI to increase efficiency, innovation and speed to market while ensuring that this technology is used in a pragmatic, secure, and responsible way.

In this edition of the Capco Journal, we are excited to share the expert insights of distinguished contributors across academia and the financial services industry, in addition to drawing on the practical experiences from Capco's industry, consulting, and technology SMEs.

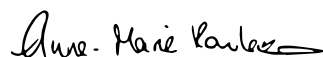
The authors in this edition offer fresh perspectives on the mindful use of GenAI and the implications of advanced GenAI on financial markets, in addition to providing practical and safe frameworks for boards and firms on how to approach GenAI governance.

The latest advancements in this rapidly evolving space demonstrate that the potential of GenAI goes beyond automating and augmenting tasks, to truly helping organizations redefine their business models, processes and workforce strategies. To unlock these benefits of GenAI, I believe that firms need a culture that encourages responsible experimentation and continuous learning across their organization, while assessing the impact of the potential benefits against a strategic approach and GenAI framework.

I am proud that Capco today remains committed to our culture of entrepreneurialism and innovation, harnessed in the foundation of our domain expertise across our global teams. I am proud that we remain committed to our mission to actively push boundaries, championing the ideas that are shaping the future of our industry, and making a genuine difference for our clients and customers – all while ensuring to lead with a strategy that puts sustained growth, integrity and security at the forefront of what we do.

I hope you'll find the articles in this edition both thought-provoking and valuable as you create your organization's GenAI strategy and future direction. As we navigate this journey together, now is the time to be bold, think big, and explore the possibilities.

My greatest thanks and appreciation to our contributors, readers, clients, and teams.



Annie Rowland, **Capco CEO**



TECHNOLOGY

08 Mindful use of AI: A practical approach

Magnus Westerlund, Principal Lecturer in Information Technology and Director of the Laboratory for Trustworthy AI, Arcada University of Applied Sciences, Helsinki, Finland

Elisabeth Hildt, Affiliated Professor, Arcada University of Applied Sciences, Helsinki, Finland, and Professor of Philosophy and Director of the Center for the Study of Ethics in the Professions, Illinois Institute of Technology, Chicago, USA

Apostolos C. Tsolakis, Senior Project Manager, Q-PLAN International Advisors PC, Thessaloniki, Greece

Roberto V. Zicari, Affiliated Professor, Arcada University of Applied Sciences, Helsinki, Finland

14 Understanding the implications of advanced AI on financial markets

Michael P. Wellman, Lynn A. Conway Collegiate Professor of Computer Science and Engineering University of Michigan, Ann Arbor

20 Auditing GenAI systems: Ensuring responsible deployment

David S. Krause, Emeritus Associate Professor of Finance, Marquette University

Eric P. Krause, PhD Candidate – Accounting, Bentley University

28 Innovating with intelligence: Open-source Large Language Models for secure system transformation

Gerhardt Scriven, Executive Director, Capco

Tony Moenicke, Senior Consultant, Capco

Sebastian Ehrig, Senior Consultant, Capco

38 Multimodal artificial intelligence: Creating strategic value from data diversity

Cristián Bravo, Professor, Canada Research Chair in Banking and Insurance Analytics, Department of Statistical and Actuarial Sciences, Western University

46 GenAI and robotics: Reshaping the future of work and leadership

Natalie A. Pierce, Partner and Chair of the Employment and Labor Group, Gunderson Dettmer

MINDFUL USE OF AI: A PRACTICAL APPROACH¹

MAGNUS WESTERLUND | Principal Lecturer in Information Technology and Director of the Laboratory for Trustworthy AI, Arcada University of Applied Sciences, Helsinki, Finland

ELISABETH HILDT | Affiliated Professor, Arcada University of Applied Sciences, Helsinki, Finland, and Professor of Philosophy and Director of the Center for the Study of Ethics in the Professions, Illinois Institute of Technology, Chicago, USA

APOSTOLOS C. TSOLAKIS | Senior Project Manager, Q-PLAN International Advisors PC, Thessaloniki, Greece

ROBERTO V. ZICARI | Affiliated Professor, Arcada University of Applied Sciences, Helsinki, Finland

ABSTRACT

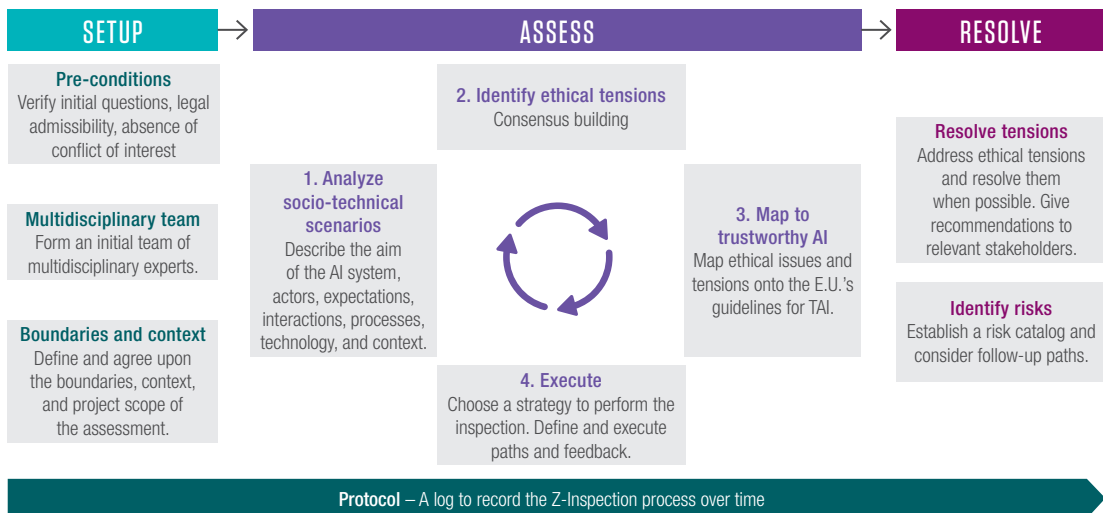
The current landscape of assuring AI reliability and quality is fragmented, with existing frameworks often lacking a unified methodology for comprehensive evaluation, particularly in integrating ethical and human rights considerations. This article introduces the Z-Inspection[®] process as a participatory, human-centered approach for assessing and co-designing trustworthy AI systems throughout their lifecycle. By forming multi-disciplinary teams and utilizing socio-technical scenarios, Z-Inspection[®] enables the exploration of ethical dilemmas and risks in context, fostering a shared understanding among stakeholders. This methodology aligns with the European AI Act's emphasis on human-centric technology and addresses limitations in existing standards by incorporating continuous ethical reflection and adaptability. We demonstrate how the co-design aspect of Z-Inspection[®] facilitates proactive risk identification, transparency, and alignment with regulatory requirements. This approach advances beyond traditional static checklists, offering a dynamic framework that intrinsically weaves ethical considerations into AI development, thereby ensuring that AI technologies are not only technically robust but also ethically sound, socially beneficial, aligned with human values, and legally compliant. Trustworthy AI is not an afterthought or technical hindrance but a way to promote a mindful use of AI.

1. INTRODUCTION

A fragmented approach to assessment and implementation characterizes the current landscape of assuring AI reliability and quality. Existing frameworks, such as the E.C. High-Level Expert Group on Artificial Intelligence (AI HLEG), National Institute of Standards and Technology (NIST), OECD, and Google's AI principles, provide valuable insights but lack a unified methodology for comprehensive evaluation. The work on standards (ISO, IEEE) and the CEN/CENELEC (cencenelec.eu) harmonized E.U. standards struggle to incorporate ethical and human rights aspects into the compliance and audit process. Particularly for high-risk AI systems in the public sector, conformity assessments are

unlikely sufficient to determine the risk to human rights and ethics when considering AI systems such as generative AI (GenAI). Concept-based assessments focusing on individual aspects like accountability, fairness, and explainability have been developed, but they often operate in isolation, failing to capture the interdependencies of these elements. Industry-specific frameworks (e.g., IEEE P2247.4) and human rights-based approaches [Dutch Fundamental Rights and Algorithm Impact Assessment (FRAIA), Council of Europe Framework Convention on AI]) have emerged, but their integration into cohesive, widely applicable standards remains a challenge. Current explainable AI (XAI) solutions, while advancing rapidly, still struggle to balance robustness and efficiency with user-friendly interpretability, especially in complex

¹ This work was co-funded by the European Union under GA no. 101135782. Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or CNECT. Neither the European Union nor the granting authority can be held responsible for them.

Figure 1: Z-Inspection® process for trustworthy AI

Z-Inspection® process flow describing the main steps of the setup, assess, and resolve phases. In parallel to the phases, a log is kept in which the process and events of the assessment are tracked. Adapted from Zicari et al. (2021b)²

domains like automated systems and GenAI. Moreover, the ethical implications and human rights considerations in AI development and deployment are often treated as secondary concerns, rather than being intrinsically woven into the fabric of AI systems from conception to implementation.

The Z-Inspection®³ process for trustworthy AI (Figure 1) offers a different path to assessing AI trustworthiness throughout the AI system lifecycle.⁴ Z-Inspection® is a validated participatory process based on human expertise that follows the AI HLEG requirements and breaks them down to deliver an ethical understanding of issues regarding specific AI use.⁵ The Z-Inspection® process can be applied to the entire AI lifecycle, typically including (1) design, (2) development, (3) deployment, (4) monitoring, and (5) decommissioning.

Recent Z-Inspection® work includes a study in collaboration with the Dutch government to combine the trustworthy AI assessment with an FRAIA⁶ that was accomplished with great success. The work highlights the importance of capturing future intentions early. It also emphasizes considering how people may be affected, by developing socio-technical scenarios that consider the broader contextual use of AI technology. This helps to avoid, for example, system-of-

systems issues when model output propagates. Such issues are complex to capture with a product-centric regulation or standard and demand a broader discussion.

The same approach can also be employed to co-design trustworthy AI systems. The socio-technical scenarios can be developed early on, during the requirements elicitation, together with the technology providers implementing the AI system. Key insights can help define a more complete set of non-functional system requirements while also guiding the core functionalities of the system, i.e., the functional requirements, towards a system architecture that is more likely to deliver trustworthy results.

2. EUROPEAN AI ACT

The European Commission has introduced a regulation that wants to “ensure a consistent and high level of protection of public interests as regards health, safety and fundamental rights” (AI Act, recital 7). The ambition is that all deployed AI systems in the E.U, are based on human-centric technology, with the ultimate aim of AI increasing human wellbeing, especially considering the risk level of an AI system (Figure 2).

² Vetter, D., et al., 2023, “Lessons learned from assessing trustworthy AI in practice,” *Digital Society* 2:3, 35

³ Z-Inspection is a registered trademark distributed under the terms and conditions of the Creative Commons (Attribution-NonCommercial-ShareAlike CC BY-NC-SA) license (z-inspection.org)

⁴ Zicari, R. V., et al., 2021, “Z-Inspection: a process to assess trustworthy AI,” *IEEE Transactions on Technology and Society* 2:2, 83-97

⁵ Allahabadi, H., et al., 2022, “Assessing trustworthy AI in times of COVID-19: deep learning for predicting a multiregional score conveying the degree of lung compromise in COVID-19 patients,” *IEEE Transactions on Technology and Society* 3:4, 272-289

⁶ Gerards, J., M. T. Schäfer, I. Muis, and A. Vankan, 2021, “Fundamental rights and algorithms impact assessment (FRAIA),” *Rijksoverheid*, <https://tinyurl.com/y75hfh5s>

The legislation is influenced by the definition of trustworthy AI (TAI), and by enacting the regulation, the Commission considers it a key aspect of Europe being a leader in TAI solutions. TAI was defined by the Commission's appointed High-Level Expert Group on Artificial Intelligence in 2019 and is based on four "ethical principles" – (1) respect for human autonomy, (2) prevention of harm, (3) fairness, and (4) explicability – and seven "requirements" that are closely related to these principles: (1) human agency and oversight, (2) technical robustness and safety, (3) privacy and data governance, (4) transparency, (5) diversity, non-discrimination and fairness, (6) societal and environmental wellbeing, and (7) accountability. The ethical principles are considered imperatives that AI practitioners should always adhere to. However, the HLEG already foresaw that the situation may arise where there are tensions between the principles and that new requirements will emerge as the technology develops and the use of AI becomes more integrated. The HLEG developed an initial checklist for practitioners to consider, but as the field has evolved, this checklist can no longer be considered complete. Furthermore, the limits of using predetermined checklists are that they are usually not dynamic enough to capture ethical reasoning.

3. HARMONIZED STANDARDS

By crafting harmonized E.U. standards that organizations can use to certify their solutions, the E.U. hopes to make the implementation of the AI Act easier than is the case with, for example, the General Data Protection Regulation (GDPR). The work for harmonized standards was given to the CEN/

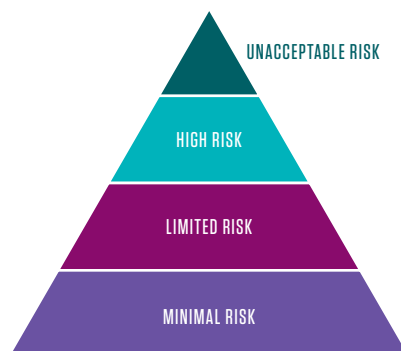
CENELEC standardization body and has yet to be completed. Harmonized standards will be created in collaboration with other international standardization bodies. However, there are some specific legal mandates that require new perspectives. One such standard is the conformity assessment standard, which should define the scope of what companies should deliver to ensure compliance.

The current preparation for a conformity assessment technical report has revealed that the requirements for ethical concerns are not directly part of what CEN/CENELEC can deliver. A fundamental difficulty in assessing ethical concerns for the purpose of a certification is that responses are not binary (pass/not pass) but may present dilemmas or a spectrum of voices that require further exploration. Treating the AI system as an isolated component makes it easier to audit technical conformity. However, the new E.U. legislation demands that the resulting AI system is trusted and trustworthy, and the problem lies herein. A company may receive a certification for a model, but integrating the model into a more extensive pipeline and the continuous operation of this system in a particular context is a very different problem than presented by the individual model. In fact, it has been shown by the MIT AI Risk Repository that most risks (65%) emerge after the AI system has been deployed.⁷ Thus, we must ask, what is the value of certification if ethical or societal concerns are not addressed in the context of applying the AI, and if continuous use modifies the data, model, or pipeline, or, even more concerning, depends on a secondary model?

4. THE Z-INSPECTION® SELF-ASSESSMENT PROCESS FOR TRUSTWORTHY AI

Our work within the Z-Inspection® initiative has taken a different approach that aims to establish a self-assessment process for AI practitioners and procurement teams that want to evaluate an AI system/component in a real-world environment. The process is participatory and seeks to consider the AI HLEG principles and requirements by forming a representative multi-disciplinary team that covers each needed area. Following a structured approach, the assessment team develops an understanding of the use case, environment, and technology that allows them to project socio-technical scenarios. By using scenarios, the work allows for an exploration of past, present, and future considerations. The team then uses a meta-framework for the claim-arguments-evidence (CAE) analysis⁸ of what was discovered to establish which claims are actual

Figure 2: The four-level risk-based approach defined within the AI Act



Adapted from: <https://tinyurl.com/bcjsjkd9>

⁷ Slattery, P., et al., 2024, "The AI Risk Repository: a comprehensive meta-review, database, and taxonomy of risks from artificial intelligence," arXiv preprint arXiv:2408.12622

⁸ Bloomfield, R., and J. Rushby, 2020, "Assurance 2.0: A manifesto," arXiv preprint arXiv:2004.10474

and which are not. Our assessment experience is that this is a very fruitful stage to establish a shared understanding of intentions and to limit the introduction of risks going forward.⁹

Following the establishment of actual claims, the process determines what evidence exists to support such claims. This work is usually done in a domain-specific manner by experts to allow for an in-depth study of concerns. Examples of such experts can be technical (machine learning and/or software architecture), legal, ethics, human rights, and, of course, domain experts from the actual environment where the system will be deployed, such as medical doctors, ecologists, and economists. Once the evidence has been gathered, it is shared among the entire group of experts, who can then still revisit their own conclusions. The final step is to verify the presented arguments that link the claims and evidence.

Based on the CAE review, an intermediary report is created and presented to the case owner, and the discussion is then aimed at resolving outstanding concerns. There are often situations that require ethics expertise, particularly to find and classify dilemmas as true or false. A vital aspect of the process is not to act as an authority that sits above the practitioners but rather as a council of peers that first helps define the solution, establish scenarios to reason within, and finally provide an outside view of what evidence is present that can validate the claims. Here, it is essential that the self-assessment team is constructed openly, without any competing interests or fear of retribution. Hence, optimally the developers themselves should not be part of the assembled expert team as they would have competing interests.

5. Z-INSPECTION® AS A TAI CO-DESIGN PROCESS

The Z-Inspection® methodology goes beyond many other TAI assessment frameworks by incorporating co-design principles throughout the AI system lifecycle.¹⁰ This co-design approach is fundamental to addressing the complex, interdisciplinary challenges posed by AI systems, particularly in high-risk domains. Integrating human-centered TAI design principles into the development work ensures that the resulting framework is not only technically robust but also accessible and meaningful to end-users and practitioners.

By facilitating a co-design process, diverse stakeholders can work together from the early stages of AI system design. This interdisciplinary collaboration ensures that multiple perspectives are considered, from conception and requirements handling to development and systems testing. By utilizing a TAI co-design process, it leads to a more comprehensive understanding of potential impacts and risks, enabling the design of a more robust, reliable, and resilient system architecture.

5.1 Co-design use case example

The co-design aspect of Z-Inspection® promotes an iterative approach to AI system development. Rather than treating ethical and societal considerations as an afterthought or a one-time compliance check, the process encourages continuous evaluation and refinement throughout the AI lifecycle. An example of Z-Inspection® functioning as a co-design process can be found in the study “Co-design of a trustworthy AI system in healthcare: deep learning based skin lesion classifier.”¹¹

In this study, the co-design methodology was applied during the early design phase of an AI system intended to assist dermatologists in diagnosing skin lesions using deep learning algorithms. For the case study, dermatologists, evidence-based medicine experts, ethicists, and patient representatives were brought together with AI engineers. This diverse group identified ethical aspects and tensions between different viewpoints, such as the varying perspectives on overdiagnosis, early detection, and prognosis-based forecasting, which might have been overlooked in a traditional development process. This interdisciplinary input helped the group of researchers and practitioners developing the tool to shape and refine the design process.

5.2 Implementation of a TAI co-design process

The agile approach yields multiple advantages throughout the AI development lifecycle. It enables proactive risk identification, ensuring that potential issues are detected and mitigated before they become entrenched in the system architecture. The methodology facilitates real-time feedback on the AI system design, empowering development teams to iteratively enhance and refine the product based on continuous stakeholder involvement. Moreover, it fosters ongoing

⁹ Vetter et al. (2023)

¹⁰ Zicari, R. V., et al., 2021, “Co-design of a trustworthy AI system in healthcare: deep learning based skin lesion classifier,” *Frontiers in Human Dynamics* 3, 688152

¹¹ Ibid.

alignment with social responsibility initiatives and evolving market expectations, ensuring the AI system maintains its relevance and ethical standing. Finally, this iterative framework cultivates organizational adaptability, reducing unforeseen AI reputational risks throughout the operational lifespan.

5.2.1 CO-DESIGN SETUP

The co-design approach starts with assembling a multi-disciplinary team comprising, for example, of AI engineers, domain experts, ethicists, legal experts, end-user representatives, and social scientists. The expert group works collaboratively to understand the AI system's aim, consider its potential impacts, and identify stakeholders' needs and concerns. Including various experts is crucial in i) understanding ethical, legal, and technical issues that could arise from the system's use, ii) assessing risks and harms, and iii) ensuring fairness.

An essential part of the setup is also to clearly define the scope of the project (including the boundaries and context of the assessment) and to create a detailed log of what is discussed and agreed to. This log will help to avoid scope creep, which often occurs in similar projects. This suggests that the team should understand the intended context and use of the AI system sufficiently to be able to, for example, analyze potential dual-use issues (unintended use of the AI).

5.2.2 SOCIO-TECHNICAL SCENARIOS

Similarly to the one-off assessment, the co-design approach uses socio-technical scenarios to establish a shared understanding of motivations and claims. Socio-technical scenarios involve the societal and technical context in which an AI system is (expected) to be used. This broad perspective avoids a narrow view in which only the tool itself and its technical aspects are assessed. These scenarios serve as a participatory design tool, enabling stakeholders to envision and explore various potential uses and impacts of the AI system in real-world contexts.

During the initial design phase of the AI system, we can start by defining TAI-related non-functional requirements and analyzing the technical functional requirements. In a current case study, an E.U. funded Horizon Europe project, "MANOLO" (GA 101135782),¹² we have employed this method to understand the AI components and system architecture that the project will deliver. In addition to requirements handling,

as this is a project that starts with a low technology readiness level (TRL), we included comprehensive desk research to proactively identify potential dependencies and consequences that may later become concerns or dilemmas. Through this approach, we hope to bridge the gap between technical capabilities and practical applications.

5.2.3 IDENTIFYING ETHICAL ISSUES AND TENSIONS

To identify ethical issues and tensions when co-designing a trustworthy AI system, we convene the multi-disciplinary team of experts and thoroughly review the proposed AI system, its intended use, and potential impacts. Depending on the project phase, the information in terms of claims, arguments, and evidence will be more or less detailed. The experts use the information that is currently available to conduct one or more structured brainstorming session(s) to surface potential ethical, legal, and technical concerns from different perspectives, considering the impacts on various stakeholders. The outcome can then be categorized and prioritized, and the identified issues are delimited into common categories like privacy and data protection, fairness and non-discrimination, transparency and explicability, safety and robustness, human agency and oversight, and accountability. Here, the ethical principles and requirements delineated in the Ethics Guidelines for Trustworthy AI serve as guidance. In a process that serves to formalize the findings, the identified ethical issues are brought in line with and mapped to the ethical principles and requirements of the European guidelines document.

The second part of this step is to analyze potential tensions between different ethical principles or stakeholder needs, such as privacy versus model performance, or explainability versus accuracy. Each identified issue and tension are documented in detail, including rationale and potential implications. To validate findings, a consensus-seeking discussion takes over to ensure the correctness both in terms of the project (scope and intentions) and also as a consensus of experts regarding the detailed issues. Finally, mitigation strategies are sought for high-priority issues, considering technical, operational, and governance approaches.

Going forward, for each new development phase, we revisit and update the ethical analysis regularly as the design and stakeholder demands evolve. This structured approach allows for the comprehensive identification of ethical concerns from multiple perspectives early in the design process. It follows along with the implementation process to ensure that the initial

¹² <https://tinyurl.com/rvvhmr8>

envisioned outcome is delivered together with the necessary evidence to support the final assessment. In fact, due to the trustworthiness related requirements and the evidence logged along the way, the final assessment becomes much more structured, facilitating the overall process of reaching a conclusion of whether the AI system assessed is trustworthy or not.

5.2.4 PRODUCTION

When a co-design product is launched into production, we enter into a new phase of the AI lifecycle that is often outside the scope of the co-design development project. However, the alignment work during the development phases of the project means that maintenance and governance practitioners have a strong body of support, and the stochastic and non-deterministic nature of the AI system may cause fewer surprises.¹³ Hence, a core outcome of the co-design approach is that it fosters transparency in the AI development process while establishing pipelines that can generate evidence toward a long-term trustworthy outcome. Involving diverse stakeholders and encouraging open dialogue helps build trust among all parties involved, including potential end-users and the broader public.

5.2.5 RISK IDENTIFICATION, MONITORING, AND COMPLIANCE

During a complex project, it may sometimes be difficult or even impossible to mitigate and resolve every identified ethical issue. The co-design approach can also be extended to include a step that identifies risks that need to be monitored and tracked throughout the design and while in production. The detailed collection of such identified risks allows them to be analyzed and reduced later. This helps the governance work to be monitored in real-time and assists in continuous compliance assurance work, also feeding system fine-tuning and improvements.

6. CONCLUSIONS

The co-design aspect of embedding trustworthy AI by using a process such as Z-Inspection® offers a dynamic and adaptable framework for addressing the ethical challenges posed by rapidly evolving AI technologies. Unlike static checklists or rigid compliance measures, this approach allows for incorporating new ethical considerations, technological advancements, and societal shifts as they emerge.

The methodology's flexibility is particularly valuable in addressing novel ethical challenges posed by cutting-edge AI technologies and responding to changing regulatory landscapes and societal expectations. By embedding diverse perspectives and continuous ethical reflection into the fabric of AI development, Z-Inspection® represents a shift in AI ethics and governance.

This approach goes beyond traditional assessment frameworks by integrating co-design principles throughout the AI system lifecycle. It facilitates collaboration among diverse stakeholders in the organization that are otherwise often working in silos. As co-design starts at the inception of AI system design it can facilitate a reduction of tensions between the areas of expertise in the organization. This interdisciplinary cooperation improves the comprehensive understanding of potential impacts and risks, leading to an improved buy-in of AI technology within the organization.

The iterative nature of the Z-Inspection® process promotes continuous evaluation and refinement throughout the AI lifecycle, treating ethical, societal and legal considerations as integral components rather than afterthoughts. This proactive approach enables early risk identification, real-time feedback on system design, and ongoing alignment with social responsibility initiatives and market expectations.

By fostering transparency and open dialogue, the co-design approach builds trust among all parties involved, including potential end-users and the broader public. It also allows for the identification and monitoring of risks that may not be immediately resolvable, supporting ongoing governance and compliance efforts.

In essence, the co-design aspect of Z-Inspection® offers a promising path toward creating human-centric AI systems that are not only technically robust but also ethically sound, socially beneficial, and regulatory compliant. This holistic approach to AI development and assessment is crucial in ensuring that AI technologies align with human values and societal needs as they continue to advance and integrate into various aspects of our lives. This approach also strives to promote the use of innovative technology and to improve design process maturity.

¹³ Dürder, B., F. Möslin, N. Stürtz, M. Westerlund, and R. V. Zicari, 2021, "Ethical maintenance of artificial intelligence systems," in Pagani, M., and R. Champion (eds.), *Artificial Intelligence for Sustainable Value Creation*. Edward Elgar Publishing

UNDERSTANDING THE IMPLICATIONS OF ADVANCED AI ON FINANCIAL MARKETS

MICHAEL P. WELLMAN | Lynn A. Conway Collegiate Professor of Computer Science and Engineering University of Michigan, Ann Arbor

ABSTRACT

The rapid advancement of surprisingly capable AI is raising questions about AI's impact on virtually all aspects of our economy and society. The nexus of AI and finance is especially salient, building on the impact AI has already had on trading and other financial domains. New AI developments could exacerbate market manipulation, and otherwise create loopholes in regulatory regimes. Anticipating these potential impacts suggests directions for market design and policy that makes financial markets robust to advanced AI capabilities.

1. INTRODUCTION

It seems that everyone is in an excited state these days about the apparently rapid advances in artificial intelligence (AI), and its potential to solve big problems or create new ones. This excitement is warranted, on both sides. AI promises to bring us extraordinary benefits through new capabilities to expand knowledge and automate difficult tasks, and by making a variety of valuable services accessible and affordable to broad segments of our society. AI also threatens us with an array of potentially negative consequences, including risks to security posed by malicious exploitation of AI, risks to safety due to inadvertent AI behaviors, and the risk of systemic disruption to the ways we work and live. The promises and threats of AI pervade essentially every area of our economy and society, including quite distinctly the financial sector.

In this article, I focus on the nexus of AI and finance, and particularly on implications of advanced AI for financial markets. I describe at a high level how AI is employed in markets today, and some of the possible implications of the newest AI developments. Following brief background on algorithmic trading, I focus on three ways in which the latest AI technology may bring some new considerations for security, efficiency, and fairness of our capital markets.

Let us start with the necessary qualifier that the future path of advanced AI is highly uncertain. If somebody tells you they know where AI technology will be in five years or ten years – or even next year – be very skeptical. Technical breakthroughs are inherently unpredictable, and AI has a particular capacity to surprise. It has surprised us many times, most recently in Fall 2022 by ChatGPT and its ilk. Even experts with the deepest understanding of generative AI (GenAI) techniques such as large language models (LLMs) were surprised at the quality and utility of results they are able to produce. We are also sometimes surprised by limitations and weaknesses of AI technology, or roadblocks to advancement. Either way, AI is likely to keep surprising us.

Please also keep in mind that we have limited visibility into developments that are already in the pipeline. There are likely thousands of active projects aiming to harness the latest GenAI advances in novel products and services. Startup companies, corporate development teams, and public and private research labs around the world are all exploring how to put GenAI to work. Many of these will fail (or already have) but some are likely to surprise us with new capabilities and impactful use cases.

2. ALGORITHMIC TRADING

Under-the-radar development is actually the main story of deployment of AI in financial markets up to now. AI is already widely adopted in support of trading in markets, where it has had a significant impact. The shift to electronic markets over the past few decades has had many effects, notably on speed of reaction to information. One effect has been to enable implementation of algorithmic strategies developed using AI technology such as advanced machine learning. Whereas the term “algorithmic trading” does not necessarily entail that there is “AI inside”, it is surely the case that developers of trading algorithms often employ cutting-edge AI techniques. I would even go as far as to claim that algorithmic trading represents the first widespread use of autonomous agents (i.e., AI decision making without humans in the loop) in a high-stakes and economically significant domain.

Electronic markets are well suited for software agents in part due to their simple and circumscribed interfaces (data feeds and order submission through well-specified protocols), which narrows the scope of agent behaviors that must be considered. Nevertheless, it may seem surprising that financial trading would be a first domain for autonomous operation, given the stakes involved, and thus the risks. It turned out that the advantages frequently outweigh the risks. Markets place a premium on the ability to process a multiplicity of information sources at high velocity, combined with rapid response time, both of which are in the wheelhouse for algorithms. In a situation where the first to respond to information captures the profit, putting a human in the loop is simply not an option. The returns to effective strategies are such that the research and development to produce them was worth the try, and once some initial success was demonstrated, regular processes and business models could be built around them.

Gauging the exact extent and nature of AI employed in algorithmic trading today is not possible, due to a lack of public information. Trading firms do not publish information about their strategies, for obvious proprietary reasons, and they also tend to be extremely protective about information regarding broad approaches, technology employed, data and information sources, and really everything about their strategic methodology and operations. Nevertheless, there are exceptions, and some information occasionally leaks out or is inferable from hiring practices, technology investments, or market observations. As a result, we can be quite confident about the general assessment that use of cutting-edge AI for trading is pervasive in current financial markets.

The opacity of state-of-the-art trading technology is itself one source of risk. There exists a keen public interest in understanding how various trading practices affect the fairness, efficiency, and stability of financial markets. The need for open information on AI trading strategy was a major motivation for my own group’s research in this area. I should emphasize that the goal of this research – by us or others – is not to assess whether algorithmic trading in general is beneficial or harmful to financial markets. The goal of the research is to tease apart the practices and circumstances that help or hurt, and further to identify market designs or regulations that promote the beneficial practices and deter the harmful ones.

For example, we have found that algorithmic market making improves efficiency and can be beneficial to those trading for investment, particularly when markets are thin and the market makers are competitive [Wah et al. (2017)]. In thick markets, though, algorithmic market making can extract surplus from investors. Another issue that we have investigated is “latency arbitrage”: the deployment of practices that leverage miniscule advantages in response time, measured in milliseconds or microseconds, to extract profit from trades that would have happened anyway [Wah and Wellman (2016)]. We and others have advocated for a mechanism called frequent batch auctions, where markets clear at fixed intervals, such as every half-second, rather than continuously, to short-circuit the latency arms race, thus improving both fairness and efficiency [Budish et al. (2015)].

3. THE NEWEST AI

While there is still much we need to understand about today’s algorithmic trading and its effects, the latest AI developments are raising qualitatively new issues about the implications for financial markets. The pace of technical advance in AI has been quite astounding in the last decade or so, but at the risk of over-simplification let me focus on two broad categories.

- **Deep reinforcement learning (DRL):** the use of deep neural networks to represent strategies, trained using reinforcement learning [Sutton and Barto (2018)]. This approach has demonstrated enormous advances over the past decade. DRL was the technology behind Google DeepMind’s breakthroughs in the game of Go [Silver et al. (2016)] and protein folding [Jumper et al. (2021)] (recently recognized with a Nobel Prize), for example, and indeed was the basis for DeepMind’s original formation. DRL is particularly salient for algorithmic trading because

it enables the partial or full automation of strategy generation. That is, with DRL one can train a strategy that responds to market information with actions without any human expressly programming the logic of this response.

- **Large language models (LLMs):** the massive neural networks trained to generate fluent natural language responses to textual prompts [Zhao et al. (2023)]. LLMs are the technology behind chatbots (e.g., OpenAI's ChatGPT), and part of the broader category of GenAI methods that have sparked the current explosion of interest in AI. LLMs are especially significant because they open up the language channel. That is, they enable AI methods to interact with humans (or other AI systems) in natural language, as well as standard computer languages. This allows them to be deployed in situations with open-ended interfaces, not just environments explicitly crafted for programmatic interaction.

These two categories are not entirely separate. In fact, configuring LLMs to perform useful tasks requires shaping how they respond to prompts using DRL, specifically training with reward signals based on human feedback. Combining the power of massive pre-trained models with DRL is indeed one of the most promising approaches for the next generation of autonomous agents.

At a high level, both of these new AI capabilities carry the potential to dramatically extend the autonomy and scope of algorithmic trading. Automating the strategy generation process itself adds a level of autonomy, in the sense of shifting the human control to a more indirect and abstract layer of supervision. Opening the language channel enables the trading agent to act autonomously in a much broader scope of situations. In principle, a capable chatbot could trade flexibly with human securities dealers in an over-the-counter trading environment.

4. MARKET MANIPULATION

Practices that inject misleading information about market conditions can seriously compromise the transparency and thereby the fairness and efficiency of public markets. Market manipulation is an old practice, but AI may turn out to amplify the power of would-be manipulators to at achieve their manipulative purpose, with lower cost and risk of detection. In response, sophisticated machine learning techniques can also be used by market regulators for enhanced surveillance, detection, and enforcement. Measures based on machine learning, however, are subject to countermeasures that aim to

undermine or circumvent the learning system [Papernot et al. (2018)]. Use of machine learning for regulation naturally sets up what is called an adversarial learning situation [Vorobeychik and Kantarcioglu (2018)], a kind of AI arms race, between the detector and evader. An inherent feature of adversarial learning is that any advance in detection technology can be immediately exploited by the evader to improve its evasion. Where this leads in any given situation is an open question. In our market manipulation studies, we have found that evading detection also weakens the manipulation [Wang and Wellman (2020)], but whether that will always be the case we cannot be sure.

The first-order concern is that malicious parties could use AI intentionally to manipulate markets. It is also possible that AI-developed trading algorithms could produce strategies that employ manipulation or other harmful tactics, even if such manipulation was not the specified objective. In fact, our research has demonstrated the possibility of an AI independently learning to manipulate a financial benchmark, given only the objective of seeking profit [Shearer et al. (2023)]. Are current regulations regarding market manipulation adequate to handle such a situation? Much of the existing law depends on "intent" to manipulate, and how that would apply to an AI algorithm that learned manipulation on its own is unclear.

This is just one example of what I call an "AI loophole". Our existing laws, generally speaking, are written based on the assumption that it is people who make decisions. When AIs are deciders, do our laws adequately ensure accountability for those putting the AIs to work?

The second issue is specific to the advances in language processing exhibited by LLM-based systems like ChatGPT. Arguably, one of the reasons that AI has been so successful in financial trading already is that the interface to markets (streams of buy and sell orders) is so simple and circumscribed. Text processing techniques based on machine learning have also been employed in trading to some extent, but the new LLMs can potentially take this to a new level. With broad language competency, massive bodies of human-generated information become available as material that can be traded on.

The new models also provide the capacity to generate text, thus opening up new language channels for AI influence. With generative capacity, systems can actively query humans to elicit information that may not have been available otherwise.

They can also use this channel to inject misleading information, which brings us back to market manipulation. Just as human manipulators employ social media in their “pump-and-dump” schemes, we should expect efforts to amplify such messages using AI.

This manipulation concern is just a special case of the broader problem of misinformation and fraud. In the wrong hands, AI can be great technology for deploying scams. Of course, this issue is relevant well beyond the financial domain.

5. CONCENTRATION OF INFORMATION

The final issue I would like to raise here relates to how the new AI technology obtains its power through training over massive datasets. It appears that qualitative leaps in capability can come from large scale source information. A corollary is that only entities with access to such large bodies of information can produce AI systems with the greatest performance. In the realm of financial trading, this could mean that concentrations of information access and ownership could convey extraordinary advantages. This naturally raises questions about how trading on information aggregated at massive scale could affect fairness and efficiency of our financial markets.

There now exist great stores of non-public information about people and their activities that have been amassed by companies through provision of information services and other online interactions. Much of this is willfully provided to enable or improve the quality of services, and often with understanding that it may also be used for marketing or related purposes. Considerations about this are typically framed in terms of personal privacy. Protection of personal privacy is indeed an important concern, but it may be equally important to consider the strategic implications of aggregations of information, as they affect us collectively beyond the individuals associated.

6. CONCLUSION AND POLICY RECOMMENDATIONS

I have highlighted three broad issues in this article: AI loopholes, opening the language channel, and concentration of information. Each may affect the balance of power in markets, through development of “super-manipulation” capabilities, strategic advantage, or other means. These issues are just a few of the ways that new AI technologies pose novel concerns for financial markets. AI also offers the potential to protect market integrity and level the investment playing field. Which

“

Our existing laws, generally speaking, are written based on the assumption that it is people who make decisions. When AIs are deciders, do our laws adequately ensure accountability for those putting the AIs to work? ”

effects predominate will be in large part determined by how we reconsider market designs and governance mechanisms for the world of AI-powered trading. I conclude with some recommendations about some tasks we should focus on in this endeavor.

Design markets for algorithms first. In the not-distant future, it is reasonable to expect that any interaction between a person (or organization) and a market will be mediated by algorithms. Even retail investors will have access to agents that can implement execution strategies on their behalf, rather than submitting limit or market orders directly. It is thus quite plausible to design market mechanisms under the assumption that the participants are algorithms. Though the algorithms themselves may have different computational resources and information sources, they will all have access to up-to-date market state and rapid response capabilities. Accordingly, it is possible to consider principles of fairness in availability to the interface, and fine-grained control of the extent and timing of information provided.

Conduct an inventory of laws and regulations to check for AI loopholes. To the extent that conduct proscribed for human actors could evade sanction through automation by algorithms, we have a potential AI loophole. It behooves financial regulators to consider where such situations might arise (e.g., for regulations expressed in terms of intent), and patch up the legal framework to prevent such evasion. Consider what new requirements may be necessary to ensure that behavior by algorithms is traceable to accountable parties, and areas of potential misconduct that are made newly practical thanks to AI.

Build foundations for trusted information. A functional financial system depends on widespread availability of reliable information, about markets, companies, assets, and the economy in general. Misinformation, including disinformation generated and promulgated through AI, pollutes the information environment and undermines sound financial decision making. AI itself will likely not be sufficient to counter misinformation, and so we would be wise to invest in mechanisms and associated infrastructure that could positively establish foundations of trust for critical financial information.

Support development of third-party evaluation tools. Systems developed using AI methods are generally more prone to unpredictable behavior, due to their complexity and their mode of development. In particular, algorithms using models

trained via machine learning may behave in surprising ways in situations not covered by their training data. Subjecting them to rigorous testing and evaluation may increase confidence in their safety and performance. Having third parties develop the testing regimes reduces concerns about conflicts of interest, and the risk of blind spots due to coupling of design and evaluation. There are signs that the marketplace is starting to develop AI evaluation services, but this could be accelerated by development of standards and certification requirements.

Research. That an academic researcher would call for more research is perhaps the least surprising recommendation. But the truth is, there is a lot we have yet to understand about the implications of advanced AI on financial markets, especially the scope of the risks and effectiveness of preventive strategies. Creating the knowledge necessary to prepare for a financial system with AI is a compelling public interest.

REFERENCES

- Budish, E., P. Cramton, and J. Shim, 2015, "The high-frequency trading arms race: frequent batch auctions as a market design response," *Quarterly Journal of Economics* 130:4, 1547–1621
- Jumper, J., et al. 2021, "Highly accurate protein structure prediction with AlphaFold," *Nature* 596, 583–589
- Papernot, N., P. McDaniel, A. Sinha, and M. P. Wellman, 2018, "SoK: Security and privacy in machine learning," 3rd IEEE European Symposium on Security and Privacy
- Shearer, S., G. Rauterberg, and M. P. Wellman, 2023, "Learning to manipulate a financial benchmark," 4th International Conference on Artificial Intelligence in Finance, 592–600
- Silver, D., et al., 2016, "Mastering the game of Go with deep neural networks and tree search," *Nature* 529:7587, 484–489
- Sutton, R. S., and A. G. Barto, 2018, *Reinforcement learning: an introduction*, second edition, MIT Press
- Vorobeychik Y., and M. Kantarcioglu, 2018, *Adversarial machine learning*, Morgan & Claypool Publishers
- Wah, E., and M. P. Wellman, 2016, "Latency arbitrage in fragmented markets: a strategic agent-based analysis," *Algorithmic Finance* 5, 69–93
- Wah, E., M. Wright, and M. P. Wellman, 2017, "Welfare effects of market making in continuous double auctions," *Journal of Artificial Intelligence Research* 59, 613–650
- Wang, X., and M. P. Wellman, 2020, "Market manipulation: an adversarial learning framework for detection and evasion," 29th International Joint Conference on Artificial Intelligence, 4626–4632
- Zhao, W. X., et al., 2023, "A survey on large language models," <https://tinyurl.com/2eb54bbs>

AUDITING GenAI SYSTEMS: ENSURING RESPONSIBLE DEPLOYMENT

DAVID S. KRAUSE | Emeritus Associate Professor of Finance, Marquette University

ERIC P. KRAUSE | PhD Candidate – Accounting, Bentley University

ABSTRACT

The emergence of generative artificial intelligence (GenAI) systems, capable of autonomously generating diverse content, is reshaping industries while raising concerns about biases, misuse, and errors. Auditing can play a crucial role in ensuring the responsible deployment of GenAI. This discussion examines the critical importance of auditing in mitigating risks and building user confidence. Recent regulatory frameworks, such as the E.U.'s Artificial Intelligence Act and New York City's Bias Audit Law, underscore the necessity of audits for high-risk AI systems, focusing particularly on fairness and data integrity. Internally, organizations benefit significantly from conducting audits to pinpoint biases and vulnerabilities, thereby upholding ethical standards and compliance. Traditional audit firms encounter challenges due to the intricate nature and rapid advancement of AI technologies. Nevertheless, they can adapt by enhancing their expertise and collaborating closely with AI specialists. In conclusion, rigorous auditing practices are essential for navigating regulatory environments, mitigating risks, and ensuring the ethical and dependable integration of GenAI systems, fostering positive societal impact.

1. INTRODUCTION

Organizations are increasingly adopting customized generative artificial intelligence (GenAI) systems tailored to their specific requirements. This trend is significantly reshaping various industries by autonomously generating data, text, and images [Thomson Reuters Institute (2024)]. Despite its transformative potential, many managers mistakenly perceive GenAI as conventional automation rather than recognizing its role as a dynamic, supportive tool [Baier et al. (2024)]. This misunderstanding impedes the integration of GenAI's iterative learning capabilities, missing opportunities for enhancing human-AI collaboration and streamlining operations. Consequently, these organizations risk developing inadequately designed systems that could yield inaccurate or biased outcomes, thereby posing ethical concerns. Auditing has thus become essential for mitigating risks and ensuring the dependable and ethical deployment of this advanced technology [Deloitte (2023)].

As GenAI adoption expands, governments are increasingly moving to regulate its implementation [Bostoen and van der Veer (2024)]. Key examples include the European Union's Artificial Intelligence Act (E.U. AI Act) and New York City's Local Law 144, also known as the Bias Audit Law [Fuchs (2023)]. These regulations mandate audits for high-risk AI systems, focusing on critical aspects such as fairness, data integrity, and adherence to ethical and legal standards. Policymakers argue that systematic identification and resolution of potential issues through audits can enhance transparency, accountability, and public confidence in AI technologies.

This article discusses the role of auditing in ensuring the responsible development and deployment of GenAI. It explores emerging regulatory frameworks and the risks associated with unaudited systems, emphasizing the importance of both internal and external audits. Furthermore, it acknowledges the challenges auditors face in this rapidly evolving field and advocates for collaboration with AI specialists. We envision a future where such collaborative efforts lead to more effective auditing practices, paving the way for a responsible era of AI implementation.

2. UNDERSTANDING GenAI SYSTEMS

GenAI stands at the forefront of technological innovation [Accenture (2023)], driving rapid transformation across industries [Yusuf et al. (2024)]. Unlike traditional AI systems that rely on existing data for classification and prediction, GenAI possesses the unique ability to autonomously generate new data, images, text, video, and more. This capability represents a significant advancement in AI, creating unprecedented opportunities for analysis, problem solving, creativity, and personalized experiences.¹ Moreover, the emergence of internal GenAI systems holds promise in addressing previously daunting challenges and reshaping our interactions with technology [WEF (2023)].

Key characteristics of GenAI:

- **Broad applicability:** GenAI is remarkably versatile. Unlike traditional AI tools designed for specific tasks, GenAI can manage complex operations and produce diverse outputs across multiple fields [Mission Cloud (2023)]. Its applications range from conducting sophisticated research simulations to generating creative works like art and music. This flexibility makes GenAI a powerful tool for tackling a variety of challenges across different industries and applications.
- **Creativity and novelty:** GenAI surpasses the limitations of existing data, creating innovative outputs. It excels in tasks like developing storylines and persuasive writing. In drug research, GenAI can design new molecular structures, while in marketing, it customizes content to enhance customer engagement. Although GenAI can inspire human creativity with fresh ideas, it also risks narrowing creative diversity by making individuals reliant on its outputs [Doshi and Hauser (2023)].
- **Continuous learning:** GenAI systems are capable of self-improvement through iterative processes and feedback loops [Steidl et al. (2023)]. This ability to learn and adapt allows these systems to continuously refine and enhance their outputs, evolving over time.

Real-world applications of GenAI:

- **Pharmaceuticals:** GenAI is revolutionizing drug discovery by generating a wide array of molecular structures with potential therapeutic benefits. This capability allows researchers to explore numerous molecular interactions, speeding up the identification of innovative drug candidates. As a result, the development of potentially life-saving medications is accelerated [McKinsey (2024)].
- **Retail:** in the retail sector, GenAI enhances customer experiences and refines marketing strategies. By analyzing extensive consumer data, GenAI customizes product recommendations, advertisements, and pricing strategies to align with individual preferences. This personalization aims to increase customer engagement and foster loyalty [Dubois and Voll (2024)].
- **Manufacturing:** GenAI is transforming manufacturing by optimizing processes for more efficient factory operations and predicting equipment failures. It evaluates vendor quality, delivery performance, and optimizes supply chain logistics. Through simulation and predictive analytics, manufacturers can reduce costs and improve operations, thereby increasing productivity and competitiveness [Limbsiya (2023)].
- **Finance and accounting:** investment analysts use GenAI tools to evaluate market trends, assess risk, and forecast prices. Lenders leverage these tools to analyze credit histories and determine borrower creditworthiness [BCG (2023)]. Additionally, AI enhances internal audit processes by detecting patterns and anomalies in datasets, helping auditors identify risks more effectively [Kroll (2021)].
- **Insurance:** insurance companies are adopting GenAI to streamline risk assessment and policy pricing. This technology aids in setting fair prices, detecting fraud, and processing claims more efficiently. Regulations like Colorado's Algorithm and Predictive Model Governance Regulation mandate safeguards when using AI and consumer data [Colorado Division of Insurance (2023)]. These include risk assessments for racial bias, independent audits for discrimination, and reporting findings to regulators, ensuring ethical AI practices in the insurance industry [DuVarney et al. (2024)].

¹ GenAI techniques can enhance prediction models and simulations in science, enabling researchers to explore challenging scenarios, hastening discoveries, and refining models for various applications.

- **Human resources:** HR departments utilize GenAI to automate tasks such as resume screening, predicting employee performance, and developing customized training programs. New York City has implemented a law requiring independent audits for AI-powered job screening tools to address concerns about potential bias and discrimination against certain applicant groups [Weykamp (2023)].

Understanding the capabilities and applications of GenAI is essential for organizations looking to leverage its potential for innovation and growth. However, this power comes with significant responsibility [Pecan (2023)]. As GenAI increasingly integrates into various sectors of our economy and society, it is crucial to ensure its deployment is responsible and ethical.² This entails addressing ethical concerns and mitigating potential risks through thorough oversight and auditing processes.

3. RECENT AI-BASED REGULATORY REQUIREMENTS

Regulatory developments in AI highlight the importance of ensuring these systems' safety, trustworthiness, and fairness through auditing and conformity assessment processes. Two notable examples are the E.U.'s AI Act and New York City's Local Law 144, also known as the "Bias Audit Law" or "AEDT Bias Audit Law" [European Parliament (2023), New York City Council (2023)].

The E.U. AI Act, currently in the legislative process and expected to be adopted in 2025, sets out stringent requirements, particularly for high-risk AI systems. These include conformity assessments to ensure compliance with trustworthy AI principles, bias and error testing, robust governance and risk management systems, third-party audits, and standards for transparency and documentation [Simbeck (2023)].

Under the E.U. AI Act, providers of high-risk AI systems must conduct conformity assessments before introducing their systems to the E.U. market [European Commission (2024)].

These assessments are designed to evaluate the system for potential biases and errors, implement strong governance and risk management systems, and provide detailed technical documentation to demonstrate compliance. For certain high-risk AI systems, third-party audits may also be required, enhancing the credibility and objectivity of the auditing process. The overarching goal of the AI Act is to ensure that AI systems deployed in the E.U. are safe, trustworthy, and respect fundamental rights, thereby fostering public trust and confidence in AI technologies.

New York City's Local Law 144 (also known as the Bias Audit Law), was enacted in 2021, and enforcement began on July 5, 2023. It requires employers and employment agencies in NYC to comply with its regulations by conducting annual independent and impartial bias audits of any automated employment decision tools (AEDTs) they use. These bias audits evaluate whether AEDTs cause disparate impacts based on gender and race/ethnicity categories, using specific metrics such as impact ratios.³ Employers must also ensure transparency by posting a summary of the latest bias audit results on their website and notifying candidates and employees whenever an AEDT is used in employment decisions. The law is enforced by NYC's Department of Consumer and Worker Protection, with penalties for non-compliance starting at U.S.\$500 for the first violation.⁴ While it establishes a methodology to detect bias in automated scoring systems, it has been criticized for not considering the entire score distribution across diverse groups, which could detect bias more accurately [Filippi et al. (2023)].

The Colorado Division of Insurance's "Algorithm and Predictive Model Governance Regulation," effective November 14, 2023, mandates that life insurance companies using external consumer data and AI models establish a comprehensive governance framework [Colorado Division of Insurance (2021)]. This regulation aims to prevent racial discrimination by requiring measures such as AI governing principles, board oversight, employee training, internal bias risk assessments, security controls, external audits, and reporting to regulators. By targeting insurers' AI and data practices, it promotes ethical and responsible conduct in the industry, marking a significant step towards equitable AI utilization in insurance.

² There are concerns regarding unchecked AI, especially when it comes to sentient AI, as its advanced intelligence and potentially differing values could result in situations that are unpredictable and uncontrollable.

³ An impact ratio, as used in AI bias audits, compares selection rates among demographic groups to detect potential hiring discrimination in AI-generated outcomes.

⁴ Some companies are circumventing NYC's anti-bias hiring law by relocating their operations, narrowly interpreting the law, or reverting to traditional hiring methods.

The first international legally binding treaty that attempts to ensure AI systems respect human rights was adopted in 2024 [Council of Europe (2024)]. It addresses the entire AI lifecycle and seeks to establish transparency and oversight requirements. It wants all parties to adopt measures to identify, assess, prevent, and mitigate AI risks that may be incompatible with human rights standards.

Recent regulatory requirements and conventions for AI underscore the increasing importance of auditing these systems for biases, errors, and adherence to ethical and legal standards. Policymakers are pushing for rigorous auditing and conformity assessments to ensure transparency, fairness, and accountability in the development and deployment of AI. This approach not only mitigates potential risks but also bolsters public trust and confidence in AI technologies.

4. THE RISKS OF UNAUDITED GenAI SYSTEMS

The rapid advancement of GenAI calls for a careful and measured approach. AI hallucinations, where the systems produce false or misleading outputs, can occur unintentionally. These inaccuracies might stem from biases, incorrect assumptions, or limitations in the training data. The unrestricted use of this emerging technology poses serious concerns [UNESCO (2023)]. Without proper oversight, we risk facing significant financial losses, ethical issues, and cybersecurity threats.

- **Financial risks (loss of trust, regulatory scrutiny, and legal liabilities):** employing GenAI systems without rigorous auditing can expose businesses and organizations to substantial financial risks. Unethical practices or algorithmic failures can erode trust among consumers, employees, and stakeholders, leading to a loss of reputation and a subsequent decline in market share.⁵ Furthermore, industries heavily reliant on GenAI systems will face heightened regulatory scrutiny in the future. Governments and other regulatory bodies are cautious in confirming compliance with data protection and privacy laws. Non-compliance and other violations can result in substantial fines and legal liabilities, risking an organization's financial situation.

- **Ethical concerns (bias, discrimination, and unintended consequences):** another ethical concern surrounding unaudited GenAI systems is their potential to perpetuate historical bias and discrimination [Stewart (2024)]. These systems are often trained on past datasets that reflect societal and economic inequalities. Without proper auditing, they can reinforce these biases, resulting in discriminatory outcomes.⁶ For example, an employment hiring algorithm trained on historical data might inadvertently favor certain demographics, perpetuating systemic biases and limiting opportunities for qualified candidates. Even without malicious intent, using flawed or incomplete data can lead to unintended and unfair outcomes. Without rigorous auditing and oversight, these ethical concerns can result in significant discriminatory risks to individuals and society.
- **Data quality, privacy, and protection:** data is the cornerstone of GenAI systems, but without strict controls and audits, the integrity of the data used for training and inference can be compromised, resulting in inaccurate or biased outcomes [Cohen et al. (2023)]. Ensuring data quality, privacy, and protection requires a comprehensive approach. It is essential to understand data collection and processing practices, as well as storage and usage protocols, to maintain the reliability and fairness of GenAI systems. Strong privacy controls are also critical to protect sensitive information from misuse or unauthorized access. Without proper auditing procedures, there is a significant risk of compromising data integrity and potentially violating privacy regulations.
- **Cybersecurity and integrity:** in today's digital landscape, unaudited GenAI systems are highly susceptible to cybersecurity threats [Hu et al. (2021)]. Malicious actors can exploit vulnerabilities in AI algorithms or infrastructure to manipulate outcomes for financial gain, ranging from cyber extortion to creating misleading financial data and biased hiring decisions. Additionally, the integrity of GenAI systems can be compromised through data manipulation or tampering. Without audits, these systems lack the transparency needed to detect such activities, leading to unreliable and untrustworthy outputs.

⁵ A Dutch government benefits scandal, where a flawed AI algorithm falsely accused thousands of families of welfare fraud, underscores the potential for significant loss of trust in AI systems without proper safeguards [Heikkilä (2022)].

⁶ AI-based facial recognition technology can be biased due to its reliance on historical data, potentially perpetuating harm because of the sensitive data involved and its potential for unethical use [Raji et al. (2020)].

To effectively mitigate these risks, organizations must prioritize robust cybersecurity measures. This involves evaluating current security protocols and implementing rigorous output validation processes. Continuous vigilance is essential, as the cybersecurity landscape is constantly evolving and requires ongoing updates and adaptations to stay ahead of potential attacks.

To mitigate risks associated with unaudited GenAI systems, organizations should adopt comprehensive strategies. This includes implementing thorough auditing processes to assess algorithmic fairness, data integrity, and cybersecurity resilience. Emphasizing transparency and accountability in AI deployment is crucial, ensuring that stakeholders are informed about potential risks and the measures in place to address them. Collaborating with regulatory bodies and industry peers can help establish best practices and governance standards for GenAI. By proactively managing these risks, organizations can minimize potential harm and build trust in the responsible use of AI technologies.

5. THE CASE FOR INTERNAL AUDITS

Although AI currently lacks unified practices and guidelines, audits can help bridge this gap [Lam et al. (2024)]. With growing regulatory scrutiny and public attention on AI technologies, organizations are increasingly recognizing the need to perform internal audits on their GenAI systems. These audits are critical for ensuring the effectiveness of oversight, monitoring, and review mechanisms [Chan and Kim (2022)]. By conducting regular reviews of controls and processes in accordance with internal audit (e.g., the Institute of Internal Auditors' (2024) "AI auditing framework") and IT governance (e.g., COBIT [ISACA (2018a)]) frameworks and standards, organizations can proactively identify and address deviations from internal policies, ethical standards, and regulatory requirements. These regular evaluations can help ensure compliance while demonstrating a commitment to responsible AI deployment to build trust among customers, employees, and regulators.

While certain audit procedures may require the involvement of data scientists or engineers, many activities can be performed without extensive AI or machine learning skills. For example, auditors can leverage COBIT 2019 controls and activities when designing and implementing testing procedures over AI processes. A white paper issued by ISACA suggests steps for applying COBIT 2019 to the auditing of AI systems [ISACA (2018a)]. These steps include defining the strategies and objectives of the AI systems, identifying and assessing

AI-specific risks and controls, and performing testing of the identified controls. For example, internal auditors can validate that decisions reliant on AI have a traceable transaction log (i.e., audit trail) in accordance with "COBIT DSS06.05 – ensure traceability and accountability for information events".

Internal audits can also serve to detect potential biases and errors in GenAI systems. These issues can stem from biased historical training data, flawed algorithms, or inadequate validation processes. Audits provide a structured framework for examining GenAI systems to uncover biases and errors before they result in undesirable outcomes. By identifying these issues early, organizations can implement corrective measures to improve the accuracy, reliability, and fairness of their AI systems.

Overall, many of internal auditors' preexisting skills, including familiarity with risk management frameworks, critical thinking to detect and assess errors, and effective communication with programmers, data analysts, and business managers, can help organizations integrate AI processes into their business while proactively managing associated risks.

6. THE CASE FOR EXTERNAL AUDITS

As emerging technologies like GenAI continue to evolve, there is a growing need for independent external audits to verify their proper functioning. While traditional certified public accountant (CPA) firms have a strong history of ensuring financial accuracy, auditing GenAI systems requires a distinct skill set due to their complexity [Strickland (2023), Costanza-Chock et al. (2022)].

One of the primary challenges CPAs face in auditing GenAI lies in the inherent complexity of AI technology itself [Dangelo (2023)]. Unlike conventional audits focused on financial transactions and documentation, GenAI systems operate through intricate algorithms and complex data interactions. CPAs may lack the technical proficiency needed to assess the performance, fairness, and reliability of AI models. Evaluating algorithmic biases, data quality issues, and the interpretability of model outputs demands specialized expertise. Without it, CPAs may struggle to accurately identify risks and deficiencies in AI systems. Moreover, the nature of AI technology poses another significant challenge for traditional audit firms [Minkinen et al. (2022)]. Unlike static financial processes, GenAI systems continuously evolve based on new data and feedback, requiring a dynamic and iterative auditing approach. CPAs may need to develop new methodologies and tools to effectively evaluate the evolving performance and compliance of AI systems over time.

Despite these challenges, traditional audit firms have a solid foundation to build upon in auditing GenAI systems. Their experience in risk assessment, internal controls, and regulatory compliance provides a framework for evaluating the governance and oversight of AI initiatives. By leveraging existing expertise and collaborating with specialists in AI, data science, and ethics, CPAs can enhance their capabilities in auditing GenAI systems and provide valuable assurance to stakeholders.

Finally, while it is technically feasible for a CPA firm to conduct both financial and AI audits for the same client, careful consideration is essential. Upholding auditor independence is critical, particularly in financial audits for public companies subject to Sarbanes-Oxley Act restrictions. Additionally, given the absence of established legal standards in AI auditing, clear separation between audit teams is necessary. Involving AI specialists can assist in managing technical aspects, while transparent communication about potential conflicts is vital for maintaining trust and ensuring a successful audit engagement.

7. AUDITING AND AI POLICY

Key players in the technology industry are actively engaging with policymakers in Washington, D.C., to address concerns over the unchecked development of AI [Stokel-Walker (2024)]. Their strategy involves shifting the narrative from solely focusing on safety concerns to highlighting global competitiveness, particularly in response to China's advancements in AI. By framing AI as a significant economic opportunity, the technology sector aims to alleviate some lawmakers' fears about potential catastrophic scenarios [Sorkin (2024)].

In the U.S., discussions on AI regulation reflect diverse perspectives, reflecting the complex interests of technology and policy stakeholders. Initially, experts and academics warned policymakers about potential risks like AI creating lethal bioweapons or evolving to pose existential threats [Rorvig (2023)]. This led to calls for stringent regulations on advanced AI systems. However, in the absence of new federal legislation, there has been a robust lobbying effort emphasizing AI's transformative economic and societal benefits. This tension underscores the balance between regulatory caution and fostering AI innovation.

Major tech firms such as Microsoft and Meta advocate for proactive collaboration with policymakers to prioritize transparency and self-regulation in AI development [Sullivan (2024)]. They aim to establish ethical guidelines and responsible AI practices to strike a balance between innovation and risk management. However, debates continue on the most effective approach to AI governance, addressing concerns about stifling innovation while ensuring accountability and safety.

In policymaking circles, auditing emerges as a potential bridge between these conflicting perspectives. Auditing mechanisms offer a means to assess the ethical and technical aspects of AI systems, providing insights into their development, deployment, and impact [Mökander (2023)]. By mandating regular audits, policymakers can use auditing processes to inform regulatory decisions, ensuring that AI technologies adhere to ethical guidelines and safety standards. Similar to financial statement audits, these requirements could serve as common ground for stakeholders, offering a pathway to reconcile innovation concerns with the need to manage AI risks effectively. Ultimately, integrating auditing into AI regulatory frameworks has the potential to shape AI policy by promoting a balanced approach that fosters innovation while mitigating unintended consequences.

8. CONCLUSION AND FUTURE RESEARCH

Auditing plays a pivotal role in ensuring the responsible development and deployment of GenAI systems amid their transformative potential and associated risks. These risks encompass biases, ethical quandaries, and vulnerabilities in cybersecurity, necessitating proactive risk management and ethical oversight through auditing.

Recent regulatory initiatives, such as the E.U.'s Artificial Intelligence Act and New York City's Bias Audit Law, underscore the critical role of auditing in upholding compliance with ethical and legal standards. These regulations mandate audits for high-risk AI systems, focusing on aspects like fairness, data integrity, and adherence to ethical guidelines.

Unaudited GenAI systems pose diverse risks including financial losses, ethical concerns about bias and discrimination, issues with data quality and privacy, as well as cybersecurity threats. Internal audits within organizations are essential for identifying and mitigating these risks, ensuring alignment with ethical norms and regulatory mandates.

Traditional audit firms face challenges in auditing GenAI systems due to their complexity and rapid evolution. However, by expanding their expertise, adopting innovative methodologies, and collaborating closely with AI and data science specialists, these firms can effectively validate GenAI systems and bolster trust among stakeholders.

Future research directions should explore specialized auditing methodologies tailored for GenAI systems, addressing challenges such as algorithmic transparency, bias detection, and continuous learning. Additionally, research could investigate how regulatory frameworks like the E.U. AI Act and Bias Audit Law impact the design, deployment, and auditing of GenAI across different sectors. Understanding the practical implementation of ethical AI guidelines in GenAI development and auditing is crucial for balancing innovation with ethical considerations.

Moreover, developing robust collaboration models between traditional auditors and AI specialists can enhance the auditing process for complex GenAI systems. Exploring how auditing practices influence public perception and trust in AI technologies and devising strategies to enhance transparency and accountability through effective communication and reporting are vital areas of research.

In summary, organizations should prioritize auditing as a foundational component of their AI governance strategy. Integrating auditing into AI regulatory frameworks can significantly shape governmental policies on AI, fostering innovation while safeguarding against potential risks. Through rigorous audits, organizations can mitigate risks, ensure regulatory adherence, and build confidence in the responsible deployment of AI technologies.

REFERENCES

- Accenture, 2023, "What is generative AI and why is it important?" <https://tinyurl.com/4hxyh9u9>
- Baier, P., D. DeLallo, and J. J. Sviokla, 2024, "Your organization isn't designed to work with GenAI," Harvard Business Review, February 26, 1-38, <https://tinyurl.com/yzany46a>
- BCG, 2023, "How asset managers can transform with generative AI," Boston Consulting Group, July 31, <https://tinyurl.com/yc3sxf6c>
- Bostoen, F., and A. van der Veer, 2024, "Regulating competition in generative AI: a matter of trajectory, timing and tools," SSRN, <https://tinyurl.com/4dwdw5tn>
- Chan, K. K., and T. Kim, 2022, "Auditing AI governance," Internal Auditor, February 21, <https://tinyurl.com/yj45rftf>
- Cohen, I. G., T. Evgeniuc, and M. Husovec, 2023, "Navigating the new risks and regulatory challenges of GenAI," Harvard Business Review, November 20, <https://tinyurl.com/yf4f52jx>
- Colorado Division of Insurance, 2021, "SB21-169: restrict insurers' use of external consumer data and protecting consumers from unfair discrimination in insurance practices," <https://tinyurl.com/ms5bk3b7>
- Colorado Division of Insurance, 2023, "Regulation 10-1-1: governance and risk management framework requirements for life insurers' use of external consumer data and information sources, algorithms, and predictive models," <https://tinyurl.com/yr7t38b3>
- Costanza-Chock, S., I. D. Raji, and J. Buolamwini, 2022, "Who audits the auditors? Recommendations from a field scan of the algorithmic auditing ecosystem," in Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), 1571-1583, <https://tinyurl.com/bddc7j3f>
- Council of Europe, 2024, "Council of Europe adopts first international treaty on artificial intelligence," May 17, <https://tinyurl.com/4sycw3t3>
- Damiani, J., 2019, "A voice deepfake was used to scam a CEO out of \$243,000," Forbes, September 3, <https://tinyurl.com/bdzmetpy>
- Dangelo, M., 2023, "Auditing AI: the emerging battlefield of transparency and assessment," Thomson Reuters, October 25, <https://tinyurl.com/yc744mej>
- Deloitte, 2023, "Navigating the artificial intelligence frontier," <https://tinyurl.com/2et7u439>
- Doshi, A. R., and O. Hauser, 2023, "Generative artificial intelligence enhances creativity but reduces the diversity of novel content," SSRN, <https://tinyurl.com/43tvdd78>
- Dubois, J., and L. Voll, 2024, "How GenAI changes the way CPG and retail operate – and consumers too," EY, March 18, <https://tinyurl.com/y22m57wt>
- DuVarney, D., P. Schmoyer, and J. Romano, 2024, "The regulatory implications of AI and ML for the insurance industry," BakerTilly, February 19, <https://tinyurl.com/249d9uzc>
- European Commission, 2024, "Shaping Europe's digital future," <https://tinyurl.com/mr2tuj6b>
- European Parliament, 2023, "EU AI Act: first regulation on artificial intelligence," <https://tinyurl.com/mjnw6ws>
- Filippi, G., S. Zannone, A. Hilliard, and A. Koshiyama, 2023, "Local Law 144: a critical analysis of regression metrics," Cornell University, <https://tinyurl.com/3surp3wt>
- Fuchs, L., 2023, "Hired by machine: can a New York City law enforce algorithmic fairness in hiring practices?" Fordham Journal of Corporate and Financial Law 28, 185, <https://tinyurl.com/4wxex3rb8>
- Heikkilä, M., 2022, "Dutch scandal serves as a warning for Europe over risks of using algorithms," Politico, March 29, <https://tinyurl.com/22m8f2tt>
- Hu, Y., W. Kuang, Z. Qin, K. Li, J. Zhang, Y. Gao, W. Li, and K. Li, 2021, "Artificial intelligence security: threats and countermeasures," ACM Computing Surveys (CSUR) 55:1, 1-36
- ISACA, 2018a, "COBIT," <https://tinyurl.com/yckbnf3>
- ISACA, 2018b, "Auditing artificial intelligence," <https://tinyurl.com/3dp3e65m>
- Kroll, K., 2021, "Using artificial intelligence in internal audit: the future is now," Internal Audit 360, March 18, <https://tinyurl.com/mrxmhe3f>
- Lam, K., B. Lange, B. Bliil-Hamelin, J. Davidovic, S. Brown, and A. Hasan, 2024, "A framework for assurance audits of algorithmic systems," <https://tinyurl.com/33uu99bv>
- Limnasiya, J., 2023, "AI and generative AI are revolutionizing manufacturing... here's how," CIO, December 14, <https://tinyurl.com/yfyh7p7>
- McKinsey, 2024, "Generative AI in the pharmaceutical industry: moving from hype to reality," McKinsey & Co., January 9, <https://tinyurl.com/3vkt3x9>
- Minkinen, M., J. Laine, and M. Mäntymäki, 2022, "Continuous auditing of artificial intelligence: a conceptualization and assessment of tools and frameworks," Digital Society 1:3, 21
- Mission Cloud, 2023, "Internal use cases for GenAI," <https://tinyurl.com/2bwaf75>
- Mökander, J., 2023, "Auditing of AI: legal, ethical and technical approaches," Digital Society 2:3, 49
- New York City Council, 2023, "Local Law 144 of 2021: automated employment decision tools (updated)," <https://tinyurl.com/3f3r2tsz>
- Pecan, 2023, "The top 6 use cases for GenAI," <https://tinyurl.com/33w6hmmw>
- Raji, I. D., T. Gebru, M. Mitchell, J. Buolamwini, J. Lee, and E. Denton, 2020, "Saving face: investigating the ethical concerns of facial recognition auditing," in Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 145-151, <https://tinyurl.com/5fxkw44u>
- Rorvig, M., 2023, "AI is getting powerful. But can researchers make it principled?" Scientific American, April 4, <https://tinyurl.com/4v4vuwcj>
- Sorkin, A. R., 2024, "A new bid to police AI," New York Times, DealBook, May 9, <https://tinyurl.com/yy8z3va9>
- Simbeck, K., 2023, "They shall be fair, transparent, and robust: auditing learning analytics systems," AI and Ethics, 1-17
- Steidl, M., M. Felderer, and R. Ramler, 2023, "The pipeline for the continuous development of artificial intelligence models – current state of research and practice," Journal of Systems and Software 199, 111615
- Stewart, K., 2024, "The ethical dilemmas of AI," University of Southern California Annenberg School for Communication and Journalism, <https://tinyurl.com/ymc3kmuv>
- Stokel-Walker, C., 2024, "AI survey exaggerates apocalyptic risks," Scientific American, January 26, <https://tinyurl.com/cw2rt8c>
- Strickland, B., 2023, "Generative AI revolution: how auditors are leading the way," Journal of Accountancy, November 8, <https://tinyurl.com/yc7eunmw>
- Sullivan, M., 2024, "Big tech's evolving role in AI governance: shaping ethical standards," Transcend, March 21, <https://tinyurl.com/bdeayf82>
- The Institute of Internal Auditors, 2024, "The IIA's updated AI auditing framework," <https://tinyurl.com/byd3macp>
- Thomson Reuters Institute, 2024, "2024 generative AI in professional services," Thomson Reuters, <https://tinyurl.com/mrtuw8ma>
- UNESCO, 2023, "Artificial intelligence: examples of ethical dilemmas," <https://tinyurl.com/yu45n8j6>
- WEF, 2023, "Beyond the status quo: how generative AI will transform industrial operations," World Economic Forum, <https://tinyurl.com/387dxcvx>
- Weykamp, G., 2023, "New York City targets AI use in hiring: anti-bias law explained," Bloomberg Law, July 5, <https://tinyurl.com/bdf4c7rk>
- Yusuf, A., N. Pervin, and M. Román-González, 2024, "Generative AI and the future of higher education: a threat to academic integrity or reformation? Evidence from multicultural perspectives," International Journal of Educational Technology in Higher Education 21:1, 21

INNOVATING WITH INTELLIGENCE: OPEN-SOURCE LARGE LANGUAGE MODELS FOR SECURE SYSTEM TRANSFORMATION

GERHARDT SCRIVEN | Executive Director, Capco

TONY MOENICKE | Senior Consultant, Capco

SEBASTIAN EHRIG | Senior Consultant, Capco

ABSTRACT

The rapid development of Large Language Models (LLMs) has revolutionized software development, yet the predominance of closed-source models has restricted their extensive adoption. In this paper, we explore open-source Large Language Models as an alternative to closed-source models like ChatGPT, particularly for the use case of interpreting legacy software source code. We evaluate open-source models for their capacity in understanding and explaining COBOL code to a human user, a crucial task for financial institutions looking to update their legacy systems while keeping their data secure in-house.

Evaluating LLMs in this domain is challenging since there's no simple right or wrong answer to the specific types of COBOL related questions we ask. Towards this, we have benchmarked the responses obtained from various proprietary and open-source LLMs against an expert human response. This method allows us to assess which models perform best for a specific type of question and are effective in a practical context.

This article provides insights for financial institutions looking to optimize or modernize their legacy systems using LLMs as well as offering considerations for adapting and integrating these models into their IT environments.

1. INTRODUCTION

In today's rapidly evolving business landscape, the demand for efficient and versatile artificial intelligence (AI) solutions has never been higher. Large Language Models have emerged as a transformative technology and are increasingly being adopted in modern businesses to elevate customer service standards, streamline internal documentation processes, and for the creation of content in diverse knowledge domains such as marketing.

Fine-tuning Large Language Models with proprietary data and domain-specific knowledge is often the driving force behind their adoption for specific use cases. This process allows

organizations to develop highly specialized solutions optimized for their unique operational challenges. Beyond optimizing workflow automation, enhancing data analysis capabilities, or facilitating internal communication, customized LLMs serve as a versatile toolkit for driving efficiency and productivity across diverse business functions.

Furthermore, as companies integrate LLMs into their operations, the decision between deploying them through a third-party hosted service or hosting them locally gains significant importance. Hosting LLMs locally provides better control over data privacy, allows for greater customization to meet specific business needs, and can reduce operational expenses.¹

¹ <https://tinyurl.com/mrysef5w>

However, since there is a range of LLMs available, each possessing unique capabilities and performance metrics, businesses face the challenge of selecting the most suitable model for their specific needs. Moreover, given the rapid advancement in this field, there is a pressing need for methods to efficiently evaluate new models as they emerge.

We have conducted a comprehensive evaluation of various leading LLMs currently available, specifically focusing on their ability to transform COBOL source code into tailored and highly consumable knowledge nuggets. This evaluation is designed to provide a more nuanced comparative view of LLM performance for specific use cases, particularly in the context of legacy code understanding, though the methodology can be applied to other domains as well. This involves assessing how well the LLMs can respond to certain types of questions, beyond just testing their domain expertise on the topic of COBOL. We aim to benchmark not only the models' proficiency in understanding COBOL code but also their capacity to abstract and reorganize information that may be highly fragmented across the technology stack.

In this, we want to shed light on the capabilities of these LLMs in addressing real-world business tasks, including COBOL code comprehension, customer query resolution, document analysis, and content generation, by exploring their ability to interpret code within an English context. Our assessment provides practical insights into effective methodologies for testing LLMs for specific tasks, offering valuable guidance for

businesses seeking to make informed decisions regarding their LLM strategy for addressing particular business problems.

1.1 The importance of LLMs for internal data processing

What was once a question of whether to adopt LLMs has transformed into a tactical consideration of how best to integrate them into existing enterprise operational frameworks. LLMs exhibit the capacity to address client queries through chatbots, screen extensive technical documentation for specific information, and generate compelling content for platforms spanning social media, public relations, and human resources.² However, effectively unlocking these benefits demands a critical decision – whether to opt for a paid model or open source.

Paid models typically offer superior performance, yet they may require sending potentially sensitive enterprise data outside the network boundaries, which could be unacceptable for highly regulated industries like finance. While some paid models can be deployed in private mode, meaning they can be hosted and operated within the organization's internal infrastructure, it is important to also consider the associated costs. On the other hand, open-source options pose the question of whether to host locally or via a third party, adding another layer of complexity. To clarify these choices, we have created a table with some selected models outlining hosting options for paid and open-source models along with their pros and cons (see Table 1).

Table 1: Comparison of hosting options for paid and open-source models: Pros and cons

LLM MODEL	HOSTING OPTION	PAID/OPEN-SOURCE	PROS	CONS
OpenAI's GPT-o1	Proprietary Cloud	Paid	State-of-the-art natural language processing capabilities	High computational costs, limited customization options
Llama3.1	Cloud-based / On-premises	Open-source ³	Strong performance in coding and reasoning, open-source flexibility	Requires significant computational resources for deployment
WizardLM-2-8x22B	Cloud-based / On-premises	Open-source	Strong performance in coding and reasoning, open-source flexibility	Not strongly aligned particularly in terms of safety and ethical considerations
DeepSeek V2.5	Cloud-based / On-premises	Open-Source ⁴	Strong performance in coding and reasoning, open-source flexibility	Demands significant computational power
Mixtral Large 2	Cloud-based / On-premises	Paid	State-of-the-art natural language processing capabilities	Requires substantial computational resources

² <https://tinyurl.com/47epujtd>

³ <https://tinyurl.com/bddbz2ac>

⁴ <https://tinyurl.com/ynaxpm5w>

Table 2: Comparison of efficiency aspects across various deployment scenarios

DEPLOYMENT SCENARIO	SPEED	COST	REGULATORY COMPLIANCE	SCALABILITY
Local deployment	Moderate	High initial hardware procurement costs	Challenging due to regulatory requirements on hardware and data security	Limited by hardware capacity, may require additional investments for scaling
Cloud-based deployment	High	Variable based on usage and service provider	Compliance with industry standards facilitated by cloud provider certifications	Easily scalable based on cloud service offerings, pay-as-you-go model
On-premises deployment	Moderate	High initial setup and maintenance costs	Direct control over regulatory compliance measures, but requires internal expertise	Scalability limited by on-premises infrastructure, potential for costly upgrades
Hybrid deployment	Variable, depending on workload distribution	Combination of initial hardware costs and cloud service fees	Compliance challenges due to data movement between environments	Offers flexibility in scaling based on workload demands, potential cost optimization

1.1.1 CONTROL OVER DATA PRIVACY AND SECURITY

One of the primary motivations behind opting for local deployment of LLMs is the enhanced control over data privacy and security. By hosting LLMs on internal servers, companies maintain sovereignty over their sensitive information, mitigating the risks associated with third-party hosting. This approach aligns with industries governed by stringent data protection regulations, ensuring compliance and bolstering trust among stakeholders.

1.1.2 EFFICIENCY IN OPERATIONAL PROCESSES

Local deployment of LLMs brings about significant efficiencies in operational processes, particularly in data processing tasks. By leveraging the computational power of internal servers, companies can conduct intricate analyses, extract insights, and derive actionable intelligence from vast datasets in a timely manner. However, efficiency isn't solely about speed; it encompasses various other dimensions as well. For instance, if an organization needs to procure hardware to support local deployment, navigating through the procurement process, especially within regulated industries like banking, might be challenging. We have compared different efficiency aspects across various deployment scenarios in Table 2.

1.2 Assessing LLMs using COBOL code as a case study

In evaluating LLMs' performance, for example COBOL code analysis, it is crucial to understand their unique features, performance, and limitations. Quantifying these parameters aids in the selection of the most suitable model for specific needs. Establishing a repeatable process enables users to systematically evaluate both local and open-source LLMs, ensuring continuous assessment of new models against consistent benchmarks as they are released. Factors such as model size, computational requirements, and fine-tuning capabilities guide the adoption strategies. Understanding whether these models are open source or proprietary, along with their commercial availability, is essential for determining accessibility and potential integration into workflows or products. In the following sections we will provide an overview of selected LLMs and how we have evaluated their performance with respect to our COBOL code case.

2. OVERVIEW OF SELECTED OPEN-SOURCE LLMs

We explored the landscape of state-of-the-art language models as of September 2024. Although these models are available for download and analysis, not all of them may be used commercially due to licensing restrictions. Our focus narrows down to a handpicked selection of models that have demonstrated the most promising performance for understanding code.

Our assessment prioritizes two key factors: (1) the computational resources available to us, and (2) the quality of the models' outputs. Particularly, we underscore the importance of having GPUs with ample VRAM to efficiently run these models. Below is a brief overview of the models that we have used in the scope of this article.

2.1 Selection of local LLMs for evaluation

The selection of local Large Language Models for evaluation is critical for assessing their performance and capabilities across their intended usage tasks. Developers often choose specific LLMs based on factors such as model architecture, training data, and fine-tuning approaches to evaluate their effectiveness in real-world applications.⁵ Additionally, LLM leaderboards serve as valuable resources that benchmark and rank current LLMs according to different criteria and can be helpful in making an initial selection.⁶ Popular leaderboards are

for instance Big Code Models Leaderboard or LMSYS Chatbot Arena Leaderboard.^{7,8} These leaderboards enable developers to compare the strengths and weaknesses of different LLMs, guiding the selection of models for specific use cases based on their performance metrics.

2.2 Criteria for comparison

When comparing open-source Large Language Models with a focus on code-related tasks, several key criteria come into play to assess their effectiveness. These criteria include model performance, resource utilization, ease of deployment, context length, and code understanding.

1. **Model performance:** Evaluating model performance relies on benchmarks for different categories such as commonsense reasoning, reading comprehension, and code. Code benchmarks like HumanEval and MBPP test a model's ability to write Python code based on a description of the code's function, which then must pass a test.^{9,10} Another method to assess the LLM's performance involves using human evaluators to rate the responses. Experts in software development can review the quality of code generated by LLMs, providing feedback on how well the model understands and applies programming concepts, syntax, and idiomatic expressions.

Table 3: Overview of selected open-source LLMs

MODEL NAME	DESCRIPTION	PARAMETERS
DeepSeek V2.5	Combines the capabilities of DeepSeek-V2-Chat and DeepSeek-Coder-V2-Instruct, merging general conversational and coding skills.	236B
Llama-3.1-405b-instruct	405B instruct-tuned model with a 128k context window, optimized for dialogue and high performance against leading models.	405B
Llama-3.1-70b-instruct	Optimized model from Meta, fine-tuned for code-based tasks, exhibiting higher alignment with human preferences in dialogue interactions.	70B
Mixtral-8x22b-instruct	Mistral's 8x22B MoE model uses 39B active parameters out of 176B, with capabilities in math, coding and reasoning.	141B
WizardLM-2-8x22B	Microsoft AI's top Wizard model. It is an instruct fine-tune of the Mixtral 8x22B model.	141B

⁵ <https://tinyurl.com/y8yedm8j>

⁶ <https://tinyurl.com/sc735tm9>

⁷ <https://tinyurl.com/mt94a7j5>

⁸ <https://tinyurl.com/27x3eybj>

⁹ <https://tinyurl.com/2p9upajy>

¹⁰ <https://tinyurl.com/2ezh8uyc>

2. **Resource utilization:** Efficient resource utilization is essential for deploying LLMs in real-world applications. This criterion assesses how effectively the model utilizes computational resources such as CPU, GPU, memory, and storage during training and inference. Optimizing resource utilization ensures cost-effectiveness and scalability of the model.
3. **Ease of deployment:** The ease of deploying an LLM significantly affects its adoption and integration into existing software development workflows. Factors such as model size, compatibility with various programming languages and frameworks, and the availability of deployment options (like local, on-premise, or cloud-based) impact how straightforward or complex the deployment process is.
4. **Context length:** The context length refers to the number of tokens the model can effectively process and utilize in generating code-related outputs. According to OpenAI, “A helpful rule of thumb is that one token generally corresponds to ~4 characters of text for common English text. This translates to roughly $\frac{3}{4}$ of a word (so 100 tokens \approx 75 words).”¹¹ Increasing the context length allows the model to process and analyze larger amounts of data (or longer sequences of text/code) at once.
5. **Code understanding:** Code understanding is a critical aspect of evaluating an LLM’s capability in code-related tasks. This criterion assesses how well the model comprehends programming languages, syntax, semantics, and idiomatic expressions commonly used in software development. A robust code understanding capability enables the model to provide accurate and contextually relevant suggestions and completions. While direct, methodical testing of “understanding” in the human sense might not be feasible, there are indirect methods like benchmarks and human evaluation to determine how well a model has learned to interpret and generate code.

When evaluating LLMs based on these criteria, companies can make informed decisions about selecting the most suitable model for their specific applications. Table 4 below provides an overview of some key selection criteria.

2.3 Key considerations for model selection

2.3.1 LICENSING

In the context of open-source Large Language Models, it is crucial to recognize that even though models may be open-source, the code they generate could still be subject to existing licenses. The Code Llama GitHub page underscores

Table 4: Overview of some key selection criteria of LLMs

MODEL NAME	PARAMETER COUNT	CONTEXT	RAM/VRAM REQUIREMENTS IN GiB (4-BIT/8BIT/16BIT PRECISION)	MODEL SIZE IN GiB (16-BIT PRECISION)	LICENSE
DeepSeek V2.5 ¹²	236	128	118/236/472	472	DeepSeek License Agreement
Llama-3.1-405b-instruct ¹³	405	128	202.5/405/810	810	Llama 3
Llama-3.1-70b-instruct ¹³	70	128	35/70/140	140	Llama 3
Mixtral-8x22b-instruct ¹⁴	141	64	70.5/141/282	282	Apache 2.0
WizardLM-2-8x22B ¹⁵	141	64	70.5/141/282	282	Apache 2.0

¹¹ <https://tinyurl.com/5n87rj3s>

¹² <https://tinyurl.com/4mamcp2>

¹³ <https://tinyurl.com/muy74yhd>

¹⁴ <https://tinyurl.com/y589zc9n>

¹⁵ <https://tinyurl.com/mry3y4m6>

that outputs from Llama models, including Code Llama, may be governed by third-party licenses. This means that while the Llama models themselves may be open-source, the code produced using these models might incorporate third-party rights or specific licensing conditions because code segments may unintentionally mirror those with restrictive usage terms found on platforms like GitHub. Therefore, users utilizing generated code that resembles licensed programs must adhere to the licensing conditions of the original code. By understanding these nuances, developers can navigate the complexities of licensing compliance effectively and ensure the ethical and lawful use of code generated by models like Llama 2.

2.3.2 QUANTIZATION

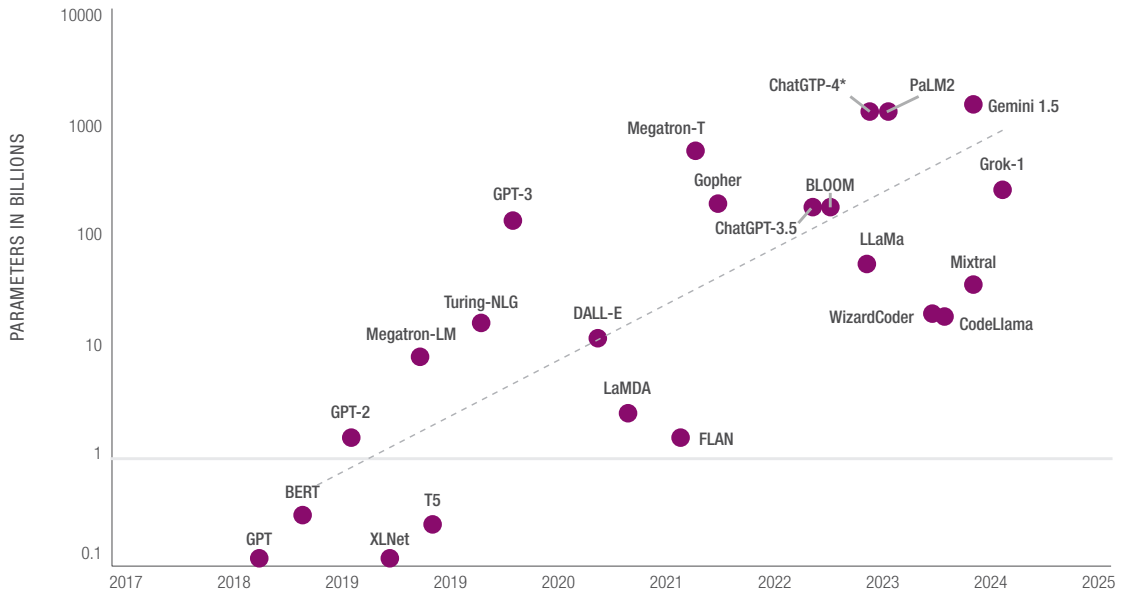
Quantization is a critical aspect to consider when selecting a Large Language Model for adoption by a company. It refers to the process of reducing the precision of numerical values in the model to enhance computational efficiency without significant loss in performance, which in turn positively affects computational resources requirements. For instance, quantizing an LLM can lead to reduced memory and computing requirements, making it more feasible for

deployment on hardware with limited resources. However, it is essential to balance these performance benefits of quantization with potential trade-offs in model accuracy. Companies should assess how different quantization techniques impact the LLM's inference speed, memory usage, and overall efficiency to ensure that the selected model aligns with their specific use case requirements and resource constraints.

2.3.3 CPU VS GPU DEPLOYMENT

When selecting an LLM, the choice between central processing unit (CPU) and graphics processing unit (GPU) for model deployment is a crucial consideration. GPUs have played a significant role in meeting the computational demands of LLMs, offering parallel processing capabilities that can accelerate model performance. Companies need to evaluate the trade-offs between CPU and GPU utilization based on factors such as performance requirements, model complexity, and available resources. While GPUs can enhance the speed and efficiency of LLM operations, they may entail higher costs and energy consumption. On the other hand, CPUs provide flexibility and cost-effectiveness but may not deliver the same level of performance for large-scale LLM tasks.

Figure 1: Evolution of Large Language Models' parameters¹⁶



¹⁶ Capco research based on model parameters from sources referenced in this paper and <https://tinyurl.com/3jx6xdwh>.
 *There are no published number of parameters available for ChatGPT4; numbers shown for ChatGPT4 are estimates according to <https://tinyurl.com/3vj7kf4r>

2.4 Evolution of Large Language Model releases

An important property of LLMs is the number of learnable elements (parameters) in a neural network, impacting their learning capacity and task performance.¹⁷ The evolution of parameter size in LLMs has seen significant growth over the years and is expected to continue for the foreseeable future (see Figure 1). This trend reflects the shift towards more complex and data-intensive models to achieve superior performance across diverse natural language processing (NLP) tasks. The increasing scale of LLMs is driven by the need for enhanced generalization, multi-modal capabilities, and improved transfer learning effectiveness. Multi-modal capabilities enable a model to comprehend various types of data, while transfer learning measures its ability to apply learned knowledge across different tasks or domains. The ongoing trend towards larger parameter sizes in LLMs underscores the continuous push towards more powerful and versatile models for advanced language understanding and generation tasks.

3. METHODOLOGY FOR COMPARISON AND EVALUATION

In the following section we will outline how we have evaluated the LLMs with respect to their ability in addressing specific tasks, within the context of their performance with COBOL code comprehension. Our evaluation process includes a dataset containing COBOL programs that are part of a COBOL application, which forms the basis for assessing the capabilities of these LLMs across different query types.

3.1 Benchmark dataset and evaluation framework

We evaluated the LLMs against a variety of tasks involving code comprehension, utilizing a diverse range of COBOL code snippets, from straightforward functions to intricate program structures, mirroring real-world scenarios commonly encountered in software development and maintenance. This evaluation was conducted using the same knowledge base for each of the models we tested.

Our evaluation framework incorporates four distinct classes of query types:

1. **Basic queries:** Evaluate the LLMs' understanding of fundamental programming concepts and COBOL code

navigation skills, such as how a function works from a technical standpoint, or where in a large piece of code a particular capability is executed.

2. **Aggregation queries:** Evaluate the LLMs' proficiency in aggregating information from various sections of the codebase, such as generating a comprehensive data dictionary. The data dictionary serves as an example of how well the model can aggregate information effectively. These queries assess the model's ability to extract and organize relevant data elements across different sections of the codebase.
3. **Reverse engineering queries:** Assess the LLMs' ability to comprehend COBOL syntax and turn it into human-interpretable forms, such as user stories, acceptance criteria, or test cases. This evaluation focuses on assessing how effectively the LLMs interpret code semantics and transform technical details into formats that are easily understandable by humans.
4. **Code improvement queries:** Evaluate the model's capability to interpret human input and suggest modifications to the code. For instance, examples include tasks like adding new data validation routines or soliciting insights on areas where code can be strengthened in response to production incidents. This evaluation focuses on assessing the models' capability to provide actionable insights for enhancing code quality and performance while preserving the integrity of the original codebase.

3.2 Benchmarking process

Each query type was submitted to the LLMs, and their responses were compared against human correct answer, i.e., answers provided by expert COBOL developers, which served as the gold standard of knowledge.

Our analysis centered on three key areas:

- Comparative analysis of open-source LLMs' performance across the query types described above
- Identification of strengths and weaknesses of each model for specific task solving
- Factors influencing model performance.

¹⁷ <https://tinyurl.com/6ekdke4a>

4. RESULTS AND CONCLUSIONS

4.1 Key findings from the comparison and evaluation

We have computed the similarity for each query to visualize and compare the performance of each model in solving specific tasks (see Table 5 below).¹⁸ We used a zero-shot approach, meaning each model was evaluated on its first attempt at answering the query. Each value in the matrix represents the cosine similarity score between the LLM responses from the model and the Human Correct Benchmark Answers, with green indicating perfect similarity and purple indicating no similarity.¹⁹ Importantly, similarity values range from 0 to 1 but do not represent accuracy percentages; rather, they indicate the degree of similarity between responses, with higher values indicating greater similarity.

The table below compares the similarity between the responses generated by different models for the four distinct query types described in Section 3.1. compared to the human provided correct answers. While the heatmap indicates some visual variability in performance across models, the overall differences in cosine similarity scores are relatively small, suggesting that most models perform at a high level in aligning with human references.

From this table, we can infer several insights regarding model performance that can guide businesses in selecting the right model for their specific use cases:

Consistency: Models that show relatively high scores across multiple query types tend to offer more consistent performance. For example, ChatGPT4o consistently delivers strong similarity scores across various tasks, making it a reliable option for broad, versatile use. Similarly, Llama3.1_405B and Gemini-pro-1.5 demonstrate steady performance, indicating adaptability across diverse queries.

Specialization: Some models excel in specific areas, making them ideal for focused use cases. For instance, Claude-Sonnet_3.5 ranks highly in the Code Update Query, highlighting its proficiency in code generation tasks. Grok-2 performs exceptionally well in the Aggregation Query, suggesting it may be the best fit for scenarios requiring data aggregation. It's important for customers to evaluate their primary use case towards making a strategic choice pertaining which LLM to select.

Table 5: Similarity between the responses generated by different models and the human-provided correct answers for various query types

MODEL	BASIC QUERY	AGGREGATION QUERY	REVERSE ENGINEERING QUERY	CODE UPDATE QUERY	AVERAGE
ChatGPT4o	High	High	High	High	High
Llama3.1_405B	High	High	High	High	High
Gemini_1.5_pro	High	High	High	High	High
Claude-Sonnet_3.5	High	Low	High	Very High	High
Llama3.1_70B	High	High	High	High	High
Grok_2	High	High	High	High	High
Wizard-LM8x22	High	Low	High	High	High
DeepSeek_v2.5	High	High	High	High	High
ChatGPT4-o1-preview	High	High	High	High	High
Mixtral_8x22B	High	High	High	High	High
Mistral_Large_v2	High	High	Low	High	High
Llama-3.2-1b-instruct	High	High	High	High	High

Perfect similarity
No similarity

¹⁸ To compute the cosine similarity between text blocks, a combined approach was used, integrating both term-frequency-based and semantic-level similarity measures.

¹⁹ <https://tinyurl.com/53jzcvss>

“

Navigating the landscape of Large Language Models demands strategic selection, where the right choice becomes the critical bridge between raw computational potential and transformative organizational intelligence. ”

Variability: Models with a wider range of similarity scores may indicate specialized strengths but could also reflect inconsistent performance. For example, WizardLM-2-8x22B shows variation across queries, excelling in some areas but performing lower in tasks like Aggregation Query, which could indicate a need to match the model with its strengths for optimal results.

Outliers: Models with lower similarity scores in certain queries highlight areas where they may struggle. However, rather than a weakness, this can be an opportunity for businesses to focus on the tasks where these models excel. For example, while Mistral-Large-v2 and Mixtral-8x22B show lower scores in Reverse Engineering, they may still be excellent choices for specific, targeted tasks if aligned with the business's key needs. Moreover, fine-tuning can significantly enhance the capabilities of even smaller models, as seen with some of the latest advances such as Llama-3.2-1b-instruct. Although this model shows lower performance across most query types, it can still be highly effective when used strategically.

4.2 Implications for companies considering the adoption of open-source LLMs for internal data processing

Performance evaluation: It's crucial for companies to conduct a comprehensive assessment of proprietary and open-source LLMs across tasks relevant to their use cases. We have selected cosine similarity as a valuable metric for comparison since it provides quantitative unbiased results.

Specialization consideration: Companies should actively design queries to comprehensively test open-source LLMs based on their specific use cases. This practical approach helps identify which models are most suitable for fulfilling their data processing needs and achieving objectives effectively.

Variability awareness: Companies should assess performance variability across different tasks or queries when adopting open-source LLMs. This includes thorough testing and evaluation of the models' capabilities across various scenarios. This assessment enables them to tailor customization or fine-tune efforts effectively, ensuring optimal performance alignment with their specific use cases.

Cost-benefit analysis: While open-source LLMs offer cost advantages compared to proprietary models, companies must weigh these benefits against potential trade-offs in performance and variability. One practical approach is to create a structured template that evaluates factors such as initial setup costs, ongoing maintenance expenses, potential productivity gains, and the expected impact on data processing efficiency.

4.3 Final thoughts on the significance of selecting the right model for specific use cases

In this article we have described a robust and repeatable method to evaluate Large Language Models across various knowledge domains, facilitating meaningful comparisons between different models.

Our strategy incorporates a flexible approach, enabling us to evaluate LLMs for any given use case. This adaptability allows us to evaluate the models efficiently within meaningful contexts as they get released, keeping pace with the latest advancements.

The framework that we have established is reusable and can be applied to many different use-cases. This structured approach systematically evaluates the costs and benefits associated with each LLM, providing stakeholders with clear insights into the value proposition and helping make informed choices that align with organizational objectives and resource constraints.

Selecting the right model for a specific use case is crucial, as it significantly impacts various aspects such as cost, footprint, and the ability to satisfy regulators. The choice of model also directly influences the performance and effectiveness of the intended use case, ensuring optimal resource allocation and compliance with regulatory standards.

5. GLOSSARY

Prompt: In the context of AI, a prompt is a text input given to a language model, which then generates an output based on the input provided.

Fine-tuning: A process in machine learning where a pre-trained model is further adjusted or 'tuned' on a new, often smaller, and more specific dataset.

Context length: The 'context length' denotes the maximum number of tokens (e.g., words, characters) an AI model can process or analyze at any given time.

Parameters: Parameters are the number of learnable parameters like weights and biases in a neural network.

Large Language Model (LLM): An LLM is a type of neural network. LLMs are typically built using neural network architectures, such as transformer models.

Transformer: The transformer model is a type of neural network architecture introduced by the landmark research paper by Google, "Attention Is All You Need", authored by eight scientists in 2017. This architecture was revolutionary for its use of self-attention mechanisms.

Token: A token refers to the smallest unit of data, usually a subword, that can be processed by the LLM.

MULTIMODAL ARTIFICIAL INTELLIGENCE: CREATING STRATEGIC VALUE FROM DATA DIVERSITY

CRISTIÁN BRAVO | Professor, Canada Research Chair in Banking and Insurance Analytics, Department of Statistical and Actuarial Sciences, Western University¹

ABSTRACT

The modern revolution of artificial intelligence (AI) has a benefit that is often not mentioned: it allows the use of diverse data from multiple sources and of multiple types (multimodal data), such as video, audio, or images, in an efficient, and, more importantly, effective manner. While this is much closer to how experts make decisions, the challenges are that it must be done profitably, while considering the internal culture and the operational systems that are available to ensure a positive return on investment (RoI). In this article, I will summarize some of the advantages and point out some of the challenges in creating effective and useful AI systems that leverage multimodal data.

1. INTRODUCTION

If you have been doing something for a long time, you probably use some sort of multimodality to make your decisions. To start this discussion, let us imagine a financial institution deciding on whether to underwrite a large bond placement in the market. An internal group of analysts will study the financial situation of the company (structured data), read reports regarding the market (text data), maybe listen to the last investor call (audio data), watch the last interview by the CEO on the local business news channel (video data), talk to experts, and analyze a long list of data sources that will help them decide if underwriting the operation is a good idea. Among these data sources, there will most likely be some scores and ratings that come from models. Maybe the analysts will use ChatGPT or other large language model (LLM) to summarize reports, but the final decision will come from interpreting all these data sources and joining them in some sort of mental or world model to arrive at a conclusion.

Hence, complex decision making is multimodal; why are our models not? This is not a capricious statement. When an expert makes a decision, it is made by combining past experiences with many sources of information in a complex, integrated manner. In these cases, AI can be a support or a hinderance. A few studies have categorically shown this. De-Arteaga et al. (2020) show that expert workers are more likely to override automated systems when the recommendations they provide go against their knowledge and experience. Lebovitz et al. (2022) find that when the AI systems that provide medical recommendations are opaque, experts will resist accepting them and fail to integrate the models into their processes. On the other hand, van den Broek et al. (2021) find that when the systems are perceived as useful, experts reach a hybrid process that enriches their knowledge with the recommendations by the AI system.

Most likely, if you work at a financial institution or a fintech company, you already have some sort of multimodal model deployed or interact with one from a provider regularly. It is common for banking apps to offer check deposits with photos,

¹ I acknowledge the support of the Canada Research Chairs Program [CRC-2018-00082].

for example. This requires the model to understand handwritten characters, connect to the structured data of the app and the customer's account information, match the information on the check with the account number, and most likely consider the response of a fraud detection model that receives the data and decides whether the deposit should be accepted or not. This can be perceived as a simple operation, but the modeling behind it is anything but. In the next couple of sections, I will highlight the advantages and challenges of deploying a multimodal model. These come from my own experience in developing these models and supporting institutions in deploying them, using, for example, text and numerical data [Stevenson et al. (2021)], LiDAR data together with sociodemographic information [Stevenson et al. (2022)], social network data in combination with behavioral data for credit risk [Zandi et al. (2024)], or combining financial information and time series market data for portfolio optimization [Korangi et al. (2024)], to name a few. Together with two colleagues, I am currently in the process of writing a book that covers the technical details of developing multimodal models in the financial services sector [Deep learning in banking, Wiley, forthcoming 2025], and I invite you to have a look if you are curious about this topic beyond this short article.

2. AI MULTIMODAL MODELS

Starting with the basics: a multimodal model is any model that uses more than one “modality” (type) of data to generate an output. A multimodal model can be a simple model that, for example, takes raw text (one, unstructured, modality) and structured information about who wrote the text (another, structured, modality), counts the number of sad emojis versus happy emojis and decides if the text has a positive or negative sentiment. Of course, this is probably a terrible model. Most modern models use some sort of deep learning AI architecture, in particular a “transformer” architecture [Vaswani et al. (2017)], to generate outputs. The transformer architecture is, in overly simple words, a statistical model that takes sequence-like data (such as text, audio, time series, and many others) and does a series of numerical processes (multiplying matrices) to generate features that describe a given outcome. This starts with generating an “embedding”, or a numerical representation of the sequence data. After what can be a massive series of calculations, the original representation is transformed into an output. This can be, for example, a probability, a forecast, or an embedding of the next word in the sequence. This latter example is the basis

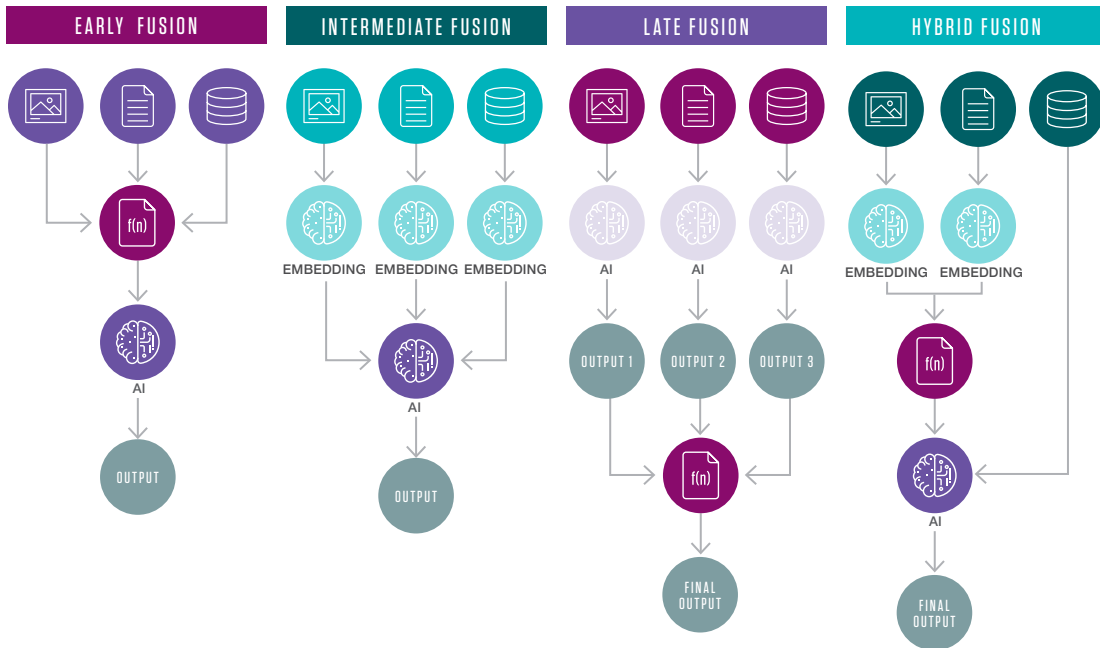
of the modern LLM systems. They take a text sequence and, through a series of transformers, predict the next word in the sequence. They do this iteratively until it predicts the sentence has reached its end.

To make a model truly multimodal, we must include many data sources. These are then combined using a data fusion strategy. We can do this at the initial step (early fusion), after it has been processed by an initial model (intermediate fusion), or even after it has been fully processed, by combining different models' outputs (late fusion). Figure 1 illustrates these fusion methods. It is also very common, and even desirable, to use more than one fusion strategy. A popular one includes using early fusion to combine similar data types, for example all numerical data modalities, and later using intermediate fusion to combine these numerical and unstructured data embeddings before processing them for prediction by dense layers. We call this process “hybrid fusion”, and it is currently the most common methodology applied in multimodal models.

The structure in Figure 1 does not showcase what exactly the embedding model, the AI model, and the output model are. This will vary depending on the application and the complexity of the model being deployed. Some examples are using featurization and then concatenation. Featurization simply means extracting numerical variables from the data, such as counting the emojis in our example above, but can become much more sophisticated. By concatenation, we mean that once the features are calculated, we create a unique numerical representation by joining the outputs of the individual features together. More sophisticated strategies use models to create these combinations. For example, in Tavakoli et al. (2023), we use Cross Attention, a type of deep learning architecture that has been shown to be useful when fusing information (J. Zhang et al. 2023). This method works by combining two or more data representations using a specialized transformer. Furthermore, lately some research has been done on using LLM themselves as the fusion strategy (Zhang et al. 2024), where now the LLM processes the output of an embedding step and tries to combine it into a fused embedding by interpreting as language.

For the embedding step, there are also many choices. The most common methodology is to use a model, tailored to that modality, that can generate a numerical output. In text, for example, transformer-based models can be used to obtain them. The package “sentence-transformer” [Ubiquitous Knowledge Processing Lab (2024)] is the best known one,

Figure 1: Different information fusion strategies



In this diagram, adapted from Tavakoli et al. (2023), there are three modalities (structured data, text, and images) and they are either processed by an aggregation function, as in the early fusion example, or they are processed by a model that outputs an embedding, a numerical representation of the data. A final AI model then transforms these modalities into a desired final output, such as a prediction or a regression value.

having a considerable library of contributed models by the likes of Google, Nvidia, and many others. For image and video data, options are less plentiful. Google has a few options in their cloud services, as part of their MediaPipe offering [Google (2024)], as does Amazon within AWS with their Titan models [Amazon AWS (2023)]. Both companies also offer direct multimodal embeddings, which use an information fusion strategy to provide a unique embedding vector that can then be fed to the output layer.

After a multimodal embedding has been created, the last step is to generate an output. This depends on the task. An AI generative model will use a series of decoders that will predict what the next embedding in a sequence is. For a prediction or regression task, it is common to create a dense neural network that generates an output in a desired format, either a probability or a regressor. These can get more exotic depending on the application. The key takeaway is that a multimodal model is a flow of smaller tasks that work together via information fusion to create a larger, complex, representation of the data. This is then fed to a final model that will construct whatever output we require.

3. CREATING A MULTIMODAL MODEL

While there are several frameworks to train a model, the objective of this article is to discuss the strategic issues and challenges that arise when planning one strategically. For the technical discussion on training these models, I refer the reader to the many online guides by vendors, to the many online resources available, or to my own previously cited works. However, deciding first if one of these models is needed is the core issue I want to start with.

3.1 Defining the problem and identifying the data

Let us start by thinking about whether there is even a need to deploy a model using multimodal data. There are three questions that a savvy manager can ask themselves:

1. Is there a problem that we have not been able to solve with traditional models?
2. Do we have data in our data lakes/data warehouses that is not being used or is underutilized?
3. What is our current technical capacity?

Figure 2: A simple project prioritization matrix

	Low Rol	High Rol
High risk	No-go	Opportunities
Low risk	Back burner	Quick wins

Quick wins are models that should take priority. Opportunities are difficult models that can create “moats”. Back burners are projects that can be developed if there is spare capacity. No-go are projects with low ROI that are also very risky to develop.

Within different organizations, there will most likely be a mix of answers to these questions. Let us imagine a traditional bank that has been collecting the social media mentions of their SME customers. This information is used in their marketing propensity models by simply counting the number of mentions in a 90-day period, to identify if the SME is generating buzz and may have a need for funds the bank can provide. This information is combined with financial transaction and a simple regression model generates a propensity probability. Starting with the first question, the answer may be that the model fails to identify negative buzz, and thus it is generating incorrect offers to companies that are in a downturn or subject to media controversy. The second question follows, and the answer is that we are indeed underutilizing the text data in the data lake as we are only generating this buzz indicator, without considering the context that a more sophisticated model can bring. The third question is much more complex to answer. If the managers of our example institution desire to move forward with a more complex multimodal model, they will need to identify if they have the correct collaborators who can develop the model in their data science teams, and whether there is on-premises or cloud infrastructure that can be used to develop the models. Cloud GPU infrastructure can get expensive fast, and on-premises infrastructure may not be sufficient to train models, only to deploy them.

With the answer to these questions, an Rol analysis must be performed to identify the data, and the training and deployment costs are balanced to merit moving forward with the model. It has been famously said that 50% of data science models fail. This is, in large part, due to not identifying which models have good Rol and which ones do not. A strategy that has worked well for my own collaborations with corporate partners has been to characterize them on a simple matrix depending on risk of development failure versus Rol, as in Figure 2.

In this prioritization, which admittedly requires significant knowledge of the organization to correctly utilize, it is easy to identify which models have the highest priority. Low risk/High Rol models (Quick wins) are the ones that should take

precedence as they are most likely easily developed, deployed, and will have a smaller risk of failure. However, these are also models that competing companies with a similar sophistication level can develop just as easily. They will most likely become commonplace very soon.

The second quadrant is far more interesting, excusing the bias that we academics have for shiny new solutions. These are far more complex models to develop, which means they cannot be easily copied by direct competitors. A successful project from this quadrant corresponds to what now is famously referred to as a “moat”. A development that gives a competitive advantage. If a company can successfully develop a model in the “opportunity” quadrant, they will have something that is challenging to replicate and provides high return. However, the risk of failure here is much higher, so strong leadership is a must.

The other two quadrants are important because projects that fit in them must be identified, and resources will most likely be better used elsewhere. It is difficult to acknowledge that an idea that one cherishes is a back burner, meaning its difficulty is low, but its return is also comparatively low against other ideas. These are the projects that can be developed whenever there is spare capacity, but if given priority, they will result in low impact and can cause more harm than good. The final quadrant, “no-go”, includes models with low potential return and high risk. These models should not be developed, better alternatives certainly exist.

How to determine Rol? This has been acknowledged as a difficult challenge [PwC (2024)]. Some factors that involve Rol can be seen in Figure 3. For the risk of development, this will mean balancing what we identified before: technical skills of the company, current computational resources available, and the ever-important cultural challenge of changing or intervening a process. The last one is one of the most significant ones in modern generative AI (GenAI). It can cause rejection within teams or customers if it is quasi-human, but clearly robotic (the uncanny valley effect), or if there is fear that it will replace jobs and should be sabotaged. This has to be managed and monitored internally, providing proper training on the use of the tool, and reassurance of its purpose within the organization.

If a model is flagged for development, then the data collection must occur. This should be done by the organization’s data engineers. Cleaning it and leaving it in a state that can be used and generating ETL pipelines that can be deployed must be accounted for when calculating the project’s Rol and risk. Next comes model training.

3.2 Training the model

At this stage, the decision made in the previous step on the strategy to train the model must be followed. The financial services sector, and most non-tech companies for that matter, have the challenge that they do not normally have the computational resources necessary to train the model, but most of the time they do have the capacity to deploy the model. Here is where scalable cloud infrastructure helps. Normally, training is far more resource-intensive than deploying a model (except for some specific high-frequency models), so a cloud solution may be best, unless the organization has significant computational resources available. Most of the time this happens, it is because the organization has mature data science teams that are constantly piloting new models. In this case, on-premises training infrastructure makes sense, as cloud costs can quickly skyrocket. This is especially true for multimodal models, where their very nature makes them far more resource-intensive than their unimodal counterparts.

Training a multimodal model is a challenge for even the most sophisticated institutions. For example, the model we developed in Korangi et al. (2024) took well over two weeks to train on a distributed system with tens of modern GPU units, and I would consider this a relatively modest multimodal problem. Cloud training of a similar model using spot Amazon

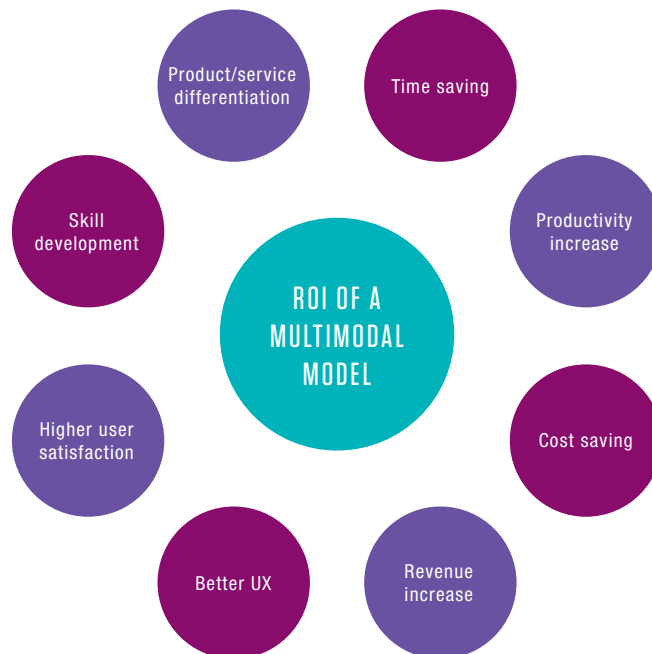
AWS instances would have run well into the six figures. Again, these calculations should have arisen from the RoI calculations in the previous step.

To effectively train the models, careful management must be followed. I can refer the interested reader to our work in this area, such as the previously referred Korangi et al. (2024) in portfolio optimization, Tiukhova et al. (2024) on credit card referral marketing, Zandi et al. (2024) and Tavakoli et al. (2023) on credit risk management, or the always rising literature discussing these deployments that can be found online. Suffice it to say, multimodality requires sufficient data science expertise internally. Subject matter experts are a requisite to train these models effectively, and whether the organization has this expertise should have been analyzed before the project began.

3.3 Deploying multimodal models

Once the model is created, the next step is to deploy it effectively. This can be done either on-premises, on cloud, or an edge deployment on the user's devices. The choice to do so must come from carefully evaluating the capacity of the company and their customers' needs, but some guidelines exist. It will depend on the type of model, the frequency it will be required, the types of devices the model will be served on, and the data security required, among other factors.

Figure 3: Some RoI factors



To begin with, the type of delivery will have an impact. I assume that some sort of large multimodal model (LMM) is being delivered, with a billion+ parameters. Most models can be “ablated”, a process to eliminate some redundant weights, to be reduced in size after training. They can also be “quantized” to reduce their size even further, by representing them in lower numerical precision. These techniques must be applied and the smallest, sufficiently performant, model must be the candidate to be deployed.

To consider if deployment must happen on premises, on the cloud, or on the device, we can start by deciding who will access the model and with which frequency. Worldwide customers frequently using a complex model at random intervals, suggests on-cloud deployment is best, as the scalable, distributed, nature of the cloud can help serve customers better. If the model is used by customers accessing their private data, our cloud choices may be limited or even forbidden by regulation, making an on-premises or edge deployment mandatory. Regarding this latter example, edge deployment applies to smaller models, but their usefulness is growing. Very recently, Microsoft developed a very aggressive 1-bit quantization that can deploy models directly on the customer’s devices [Ma et al. (2024)]. This technology is extremely recent though, but I expect it will become much more mainstream. A large range of options exist here, and the decision between edge, cloud, and local deployments must be carefully considered.

A second factor is infrastructure availability. Does your organization have sufficient capacity to deploy models internally? Servers, bandwidth, IT experts, cooling, and all the different parts of modern data centers must be available for on-premises deployment. This is most likely available in larger organizations, but not so much with startups and smaller fintech companies. Ongoing costs and growth strategies must be balanced to decide if on-premises is a sustainable strategy.

And finally, a model validation and monitoring plan must be in place to measure the performance of the model. At the pilot stage, I am a firm believer that A/B tests with carefully designed measures according to the original business plans are paramount. Research has shown that companies that use A/B tests in their operations have a higher RoI and better organizational learning strategies than the ones that do not [Koning et al. (2022)]. After deployment, standard “machine learning ops” (MLOps) procedures must be followed to ensure the model performs as expected. One specific challenge to multimodal models is that their building blocks are subject

to constant improvements as new sub-models for those modalities appear, making the “continuous improvement” leg of MLOps even more important. For example, in October 2024, the company Mistral AI released a 3-billion parameter model called Ministral [Mistral AI (2024)] that they claim is better than their 2023 7-billion-parameter model. If a local multimodal model has this older model deployed, it may make financial and statistical sense to replace it by their smaller and higher performant variant. Continuous “integration, training, delivery and monitoring”, the four core components of MLOps, are even more significant in multimodality.

4. SOME SPECIFIC CHALLENGES IN THE FINANCIAL SERVICES SECTOR

To finish this discussion, I want to touch on the specific challenges that financial institutions face when deploying multimodal models. The first challenge is a common one that may be more prevalent for multimodal data: designing data pipelines around legacy systems. Many financial institutions, particularly banks, are still running legacy systems that sometimes are sources of multimodal data, such as call center recordings or SWIFT transactional records. Designing new solutions may come with unwanted or unplanned overheads, such as creating a COBOL codebase that can create a dataset needed for a multimodal model deployed using PyTorch Lightning over a cloud server. Such technological chimeras must be identified and their benefits and costs carefully balanced.

A second point, not exclusive but certainly key in financial services, is change management and the culture of the company and its customers. If this is an internal model, how it is presented and deployed internally is a significant issue. In a recent project, in the context of an internal model to support customer service agents, we realized there was significant resistance to the use of GenAI by the organization’s collaborators as it was suspected it would lead to staff reduction. This threatened the success of the project altogether. The solution was to introduce the project by showcasing how it would help them, and how the model was simply an aid, not a replacement. This human-in-the-loop approach was practical: transformer-based generative models hallucinate, so human supervision is paramount to ensure these mistakes are caught early. Once the users realized the model was there to help them and that they remained the core owners of the workflow, their opinion shifted immediately. The model became “empowering”.

The final key issue is the regulatory hurdles that are inherent to model deployment in financial services. Tackling regulation and model management is a problem about which many articles have been written. Depending on the organizational area the model is deployed to (risk, marketing, operations, or any other), there will most likely be a series of regulatory hurdles that must be tackled to deploy the model. No matter the area the model is meant to support, model governance will be vital. One of the core aspects of modern machine learning deployment is accountability, who is responsible for the model performance, usage, and monitoring. Properly defining the responsibilities of model management, and how these models fit within the regulatory requirements that the organization is subject to, will greatly improve the chances a model is successful.

5. CONCLUSION

This article discusses multimodality and how AI can now leverage multiple sources of data. If you are reading this and work at any modern financial institution, I am sure there is some multimodal data somewhere in the organization that now just generates storage costs but can become key in multimodal development.

The core issue in generating a multimodal model is to balance RoI with the risk of development and deployment. Multimodality is tricky, requiring very complex individual parts working together in tandem. For example, a text-image-structured model can have a small 3B parameter LLM to generate sentence embeddings, a vision transformer to generate image

embeddings, a dense neural network to process the structured data, a cross-attention transformer to generate multimodal fusion, and a series of dense layers to generate a prediction. This can make the model scale to the tens of billions of parameters with relative ease, so identifying the right RoI opportunities is key. However, a correctly designed multimodal model also creates a moat: it is challenging to replicate and difficult to develop by competitors, while also generating internal skills within the data science teams that will not be common in the marketplace. All these considerations must be balanced when green-lighting a multimodal AI development.

There are both generic and specific challenges that must be considered when developing and deploying models in financial institutions. The biggest generic challenge is culture change, as AI models, particularly generative ones, can cause significant resistance. The more specific challenges within the financial services sector are transparency and accountability requirements, regulatory oversight and the risk of being a first mover in regulated models, and whether some of the multimodal data comes from legacy systems. Identifying such risks, managing them, and mitigating them can be the key to avoiding failure in an otherwise technically sound deployment.

Multimodality is the near future of AI. LLMs are already evolving into “large multimodal models”. Questioning now whether current AI and machine learning developments can be enriched by multimodality, or if new multimodal models can be created that solve previously unsolvable problems, can bring competitive advantages arising from better leveraging the diversity of data modern financial institutions and their customers create.

REFERENCES

- Amazon AWS, 2023, "Amazon Titan Image Generator, multimodal embeddings, and text models are now available in Amazon bedrock," AWS News Blog, <https://tinyurl.com/yck5pbvh>
- De-Arteaga, M., R. Fogliato, and A. Chouldechova, 2020, "A case for humans-in-the-loop: decisions in the presence of erroneous algorithmic scores," Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI '20, Association for Computing Machinery, 1-12, <https://tinyurl.com/dx6t8dah>
- Google, 2024, "Image embedding task guide | Google AI edge," <https://tinyurl.com/4f53eyvw>
- Koning, R., S. Hasan, and A. Chatterji, 2022, "Experimentation and start-up performance: evidence from A/B testing," *Management Science* 68:9, 6434-6453
- Korangi, K., C. Mues, and C. Bravo, 2024, "Large-scale time-varying portfolio optimization using graph attention networks," <https://tinyurl.com/5mhxejhe>
- Lebovitz, S., H. Lifshitz-Assaf, and N. Levina, 2022, "To engage or not to engage with AI for critical judgments: how professionals deal with opacity when using AI for medical diagnosis," *Organization Science* 33:1, 126-148
- Ma, S., et al., 2024, "The era of 1-Bit LLMs: all large language models are in 1.58 bits," <https://tinyurl.com/3d3fh7bv>
- Mistral AI, 2024, "Un Ministral, Des Ministraux," <https://tinyurl.com/2unef8ep>
- PwC, 2024, "Solving AI's ROI problem. It's not that easy," PricewaterhouseCoopers, <https://tinyurl.com/dmndjrw7>
- Stevenson, M., C. Mues, and C. Bravo, 2021, "The value of text for small business default prediction: a deep learning approach," *European Journal of Operational Research* 295:2, 758-771
- Stevenson, M., C. Mues, and C. Bravo, 2022, "Deep residential representations: using unsupervised learning to unlock elevation data for geo-demographic prediction," *ISPRS Journal of Photogrammetry and Remote Sensing* 187, 378-392
- Tavakoli, M., R. Chandra, F. Tian, and C. Bravo, 2023, "Multi-modal deep learning for credit rating prediction using text and numerical data streams," <https://tinyurl.com/393b84rt>
- Tiukhova, E., E. Penalzoza, M. Oskarsdottir, B. Baesens, M. Snoeck, and C. Bravo, 2024, "INFLECT-DGNN: influencer prediction with dynamic graph neural networks," *IEEE Access* 12, 115026 - 115041, <https://tinyurl.com/msdzu86>
- Ubiquitous Knowledge Processing Lab, 2024, "Sentence transformers," sbert.net
- van den Broek, E., A. Sergeeva, and M. Huysman, 2021, "When the machine meets the expert: an ethnography of developing AI for hiring," *MIS Quarterly* 45:3, 1557-1580
- Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, 2017, "Attention is all you need," *Advances in Neural Information Processing Systems* 30
- Zandi, S., K. Korangi, M. Óskarsdóttir, C. Mues, and C. Bravo, 2024, "Attention-based dynamic multilayer graph neural networks for loan default prediction," *European Journal of Operational Research*, September, <https://tinyurl.com/yc3p66df>
- Zhang, D., Y. Yu, J. Dong, C. Li, D. Su, C. Chu, and D. Yu, 2024, "MM-LLMs: recent advances in multimodal large language models," in Ku, L.-W., A. Martins, and V. Srikumar (eds.), Findings of the Association for Computational Linguistics ACL 2024, Association for Computational Linguistics
- Zhang, J., Y. Xie, W. Ding, and Z. Wang, 2023, "Cross on cross attention: deep fusion transformer for image captioning," *IEEE Transactions on Circuits and Systems for Video Technology* 33:8, 4257-4268

GenAI AND ROBOTICS: RESHAPING THE FUTURE OF WORK AND LEADERSHIP

NATALIE A. PIERCE | Partner and Chair of the Employment and Labor Group, Gunderson Dettmer

ABSTRACT

This article explores the transformative impact of generative AI (GenAI) and robotics on the future of work and leadership. It discusses how these technologies are revolutionizing various industries, including healthcare, finance, retail, manufacturing, and education. The synergy between GenAI and robotics is highlighted, showing potential for adaptive robotics and enhanced human-robot interaction. The article emphasizes the critical role of leadership in navigating this technological shift, addressing the need for strategic vision, resource allocation, and fostering an AI-friendly culture. It also covers the importance of workforce reskilling and the use of GenAI in learning and development. Legal considerations, including data privacy, discrimination risks, intellectual property rights, and evolving regulatory frameworks, are examined. The article concludes by discussing challenges such as ethical concerns, job displacement, and data security, while emphasizing the potential for GenAI to drive innovation and competitive advantage when balanced with human-centric values and ethical considerations.

1. INTRODUCTION – THE RISE OF GenAI AND ROBOTICS

The dawn of generative AI (GenAI) marks a pivotal moment in technological advancement, ushering in an era of unprecedented change that is reshaping industries, economies, and societies worldwide. This transformative technology, coupled with robotics, presents both extraordinary opportunities and complex challenges for businesses and their leadership. As organizations rush to harness the power of GenAI to drive innovation, enhance productivity, and gain competitive advantage, they must navigate a multifaceted landscape of technological, ethical, and legal considerations.

The potential of GenAI to revolutionize everything from product development and customer experiences to workforce training and decision making processes is immense. GenAI, in particular, represents a significant leap forward in artificial intelligence (AI) capabilities. Unlike traditional AI systems

designed to analyze and interpret existing data, GenAI can create new, original content across various mediums – text, images, audio, and even code.

The global AI market size is massive. In 2023, it was valued at over U.S.\$200 billion and was projected for a compound annual growth rate (CAGR) of 37.3% through 2030.¹ The robotics market is similarly poised for explosive growth, with a CAGR of 22.8% through 2030.² These figures underscore the rapidly growing importance of AI and robotics across various sectors, and the need for business leaders to understand the impact of technology.

However, this potential comes with significant responsibilities. Leaders must not only understand and implement these advanced technologies but also grapple with critical issues such as data privacy, algorithmic bias, intellectual property rights, and evolving regulatory frameworks. The integration of GenAI into the workplace demands a delicate balance between

¹ Horizon Grand View Research, 2024, "Global artificial intelligence market size and outlook," <https://tinyurl.com/3rkscymn>

² GlobeNewswire, 2022, "Robotics market size to cross USD 214.68 billion by 2030, growing at a CAGR of 22.8% – report by Market Research Future (MRFR)," <https://tinyurl.com/e2v3nbfe>

leveraging its capabilities for business success and ensuring ethical, legal, and socially responsible use. As we embark on this new chapter of the AI revolution, the role of forward-thinking, ethically grounded leadership has never been more crucial in shaping a future where technology enhances human potential while addressing the complex challenges it presents.

2. TRANSFORMING INDUSTRIES THROUGH GenAI

The potential applications of GenAI span virtually every industry, promising to reshape business processes, customer interactions, and product development. There are also risks, but the following highlights the potential benefits across several key sectors.

2.1 Healthcare

In healthcare, GenAI is revolutionizing drug discovery, personalized treatment plans, and medical imaging analysis. By generating and screening potential molecular structures, AI accelerates the drug discovery process, potentially bringing life-saving treatments to market faster. In personalized medicine, AI analyzes vast amounts of patient data to generate tailored treatment plans, improving patient outcomes and reducing healthcare costs.

2.2 Financial

The financial services sector is leveraging GenAI for fraud detection, personalized financial advice, and predictive analytics for market trends and risk assessment. AI-powered systems can analyze complex patterns in financial transactions, identifying potential fraud more quickly and accurately than traditional methods. Moreover, these systems can generate personalized financial advice by considering an individual's financial history, goals, and risk tolerance, providing more targeted and effective financial planning services.

2.3 Retail

Retail and e-commerce are seeing a transformation in personalized product recommendations, automated content generation, and even AI-driven product design. GenAI can analyze customer behavior and preferences to create highly targeted product recommendations, increasing sales and customer satisfaction. In content creation, AI can generate product descriptions, marketing copy, and even visual content, streamlining the process of keeping online catalogs up-to-date and engaging.

2.4 Media

The media and entertainment industry is experiencing a creative renaissance with AI-generated scripts, music, and visual effects, alongside more sophisticated content recommendation systems. AI can analyze trends and audience preferences to generate initial script ideas or musical compositions, serving as a creative springboard for human artists. In visual effects, GenAI can create realistic environments, characters, and animations, reducing production time and costs for film and television projects.

2.5 Manufacturing

Manufacturing is benefiting from GenAI's ability to optimize product design, predict maintenance needs, and streamline supply chain efficiencies. AI can generate multiple design iterations based on specific parameters, allowing engineers to explore innovative solutions more quickly. In predictive maintenance, AI analyzes sensor data to forecast potential equipment failures, reducing downtime and maintenance costs.

2.6 Education

In education, we are witnessing the rise of personalized learning experiences, automated grading systems, and AI-generated educational content. GenAI can create customized learning materials that adapt to a student's learning style and pace, making education more effective and engaging. Automated grading systems powered by AI can provide instant feedback on assignments, allowing teachers to focus more on individual student needs.

2.7 Law

Legal services are being enhanced with AI assistance in contract analysis, legal research, and predictive analytics for case outcomes. GenAI can quickly analyze vast amounts of legal documents, extracting relevant information drafting contract clauses and other content, and identifying potential issues. In legal research, AI can generate comprehensive summaries of relevant case law and statutes, significantly reducing the time lawyers spend on research tasks.

3. THE SYNERGY OF GenAI AND ROBOTICS

The combination of GenAI with robotics presents a new frontier of possibilities, further amplifying the impact on industries. This synergy is enabling adaptive robotics, where robots can

generate new movement patterns to handle unfamiliar tasks or environments. In manufacturing, for instance, robots powered by GenAI can adapt to new product designs or production processes without extensive reprogramming, increasing flexibility and efficiency on the factory floor.

The integration of GenAI is also enhancing human-robot interaction through more intuitive and efficient communication, especially in collaborative work environments. Robots can now understand and respond to natural language commands, making it easier for human workers to collaborate with their robotic counterparts. This improved interaction is particularly valuable in industries like healthcare, where robots assist in surgeries or patient care, adapting their behavior based on real-time feedback and changing conditions.

In manufacturing, the synergy of GenAI and robotics could enable highly customized, on-demand production. AI systems can generate product designs based on specific customer requirements, while robots execute these designs in real-time. This approach could revolutionize the automotive, aerospace and other industries, allowing for cost-effective mass customization of complex products.

The healthcare sector stands to see significant advancements from this synergy. Surgical robots enhanced by GenAI could adapt to unexpected situations during procedures, potentially improving patient outcomes. These AI-enhanced robots could generate new surgical approaches on the fly, considering the unique anatomy of each patient and adapting to any complications that arise during surgery. Of course, the stakes are high and these advancements must be aligned with appropriate risk-mitigation measures.

4. LEADERSHIP IN THE GenAI ERA

In this rapidly evolving landscape, the role of company leadership – from C-suite executives to middle management – is more critical than ever. Leaders must not only understand the potential of GenAI but also understand the risks and actively champion integration into business strategies and operations.

Developing a clear vision for how GenAI can drive organizational growth and competitive advantage is paramount. This involves identifying key areas where GenAI can create value, whether through cost reduction, revenue generation, or improved customer experiences. Leaders must craft a comprehensive

roadmap for AI integration that aligns with the company's overall strategic goals and includes risk mitigation measures, ensuring that AI initiatives are not siloed but integrated across all aspects of the business.

Effective resource allocation is crucial in this process. Leaders must balance the allure of short-term gains with the need for long-term transformational projects. This might involve setting up dedicated AI research and development teams, investing in data infrastructure, or forming strategic partnerships with AI technology providers.

Creating a culture that embraces AI innovation is equally important. Leaders should foster an environment that encourages experimentation and learning, where employees feel safe to engage with AI technologies and explore their potential applications and risks. This could involve setting up innovation labs, hosting hackathons, or implementing reward systems for AI-driven improvements.

Promoting cross-functional collaboration is key to driving holistic AI integration. Leaders should break down silos between technical and non-technical teams, encouraging knowledge sharing and collaborative problem solving. This approach ensures that AI solutions are not just technologically sound but also aligned with business needs and user requirements.

Leading by example is crucial in the AI era. Executives and managers should demonstrate a commitment to continuous learning and adaptation, actively engaging with AI technologies and staying informed about the latest developments in the field. This might involve attending AI conferences, participating in AI training programs, or even experimenting with AI tools in their own work.

As AI becomes more pervasive, ethical leadership takes on new importance. Leaders must champion responsible AI practices, establishing clear guidelines for AI development and use within the organization. This involves ensuring transparency in AI decision making processes, addressing potential biases, and considering the broader societal implications of AI deployment.

Engaging with stakeholders on AI-related ethical considerations is also crucial. Leaders should facilitate open dialogues with employees, customers, and the wider community about the company's use of AI, addressing concerns and building trust. This transparent approach mitigates risks and positions the company as a responsible leader in the AI revolution.

5. RESKILLING THE WORKFORCE: LEVERAGING GenAI FOR LEARNING AND DEVELOPMENT

One of the key risks of GenAI is worker displacement. Ironically, one of the most powerful applications of GenAI for business leaders is in workforce development and reskilling. As the skills gap widens due to rapid technological advancement, GenAI offers innovative solutions for large-scale, personalized learning initiatives.

AI-powered systems can analyze an employee's current skill set, role requirements, and career aspirations to create tailored learning paths. These systems can generate personalized course content, adapting to each learner's pace and preferred learning style. For technical skills development, AI can create realistic simulations and scenarios, providing hands-on practice in a safe, virtual environment. In soft skills training, AI can generate various interactive scenarios, helping employees practice communication, leadership, and problem solving skills in diverse contexts. AI can also leverage real-time language translation, which can broaden the reach and effectiveness of any reskilling program.

The ability of GenAI to provide instant, constructive feedback on assignments and assessments accelerates the learning process. AI systems can analyze an employee's performance, identify areas for improvement, and generate targeted recommendations for further study or practice. This personalized feedback loop ensures that learning is efficient and directly relevant to each employee's needs.

GenAI also enables continuous skills gap analysis, allowing leaders to stay ahead of emerging skill requirements. By analyzing industry trends, job market data, and company-specific needs, AI systems can identify emerging skills gaps and predict future skill requirements. This foresight allows leaders to proactively adjust training programs, identify high-potential employees for upskilling or reskilling initiatives, and make data-driven decisions about hiring and workforce development strategies.

In the realm of knowledge transfer, GenAI can play a crucial role in preserving and disseminating institutional knowledge. AI systems can create summaries of expert knowledge and best practices, making this valuable information more accessible across the organization. They can also facilitate more effective mentorship programs by matching employees

based on complementary skills and development needs, and even generate additional resources to support these mentorship relationships.

6. ACHIEVING BUSINESS SUCCESS IN A RAPIDLY EVOLVING LANDSCAPE

GenAI offers powerful tools for leaders to drive business success in an increasingly complex and fast-paced environment. In the realm of decision making, AI can augment leadership processes by analyzing vast amounts of data to generate insights and predictions. It can create detailed scenario models to assess potential outcomes of strategic decisions, allowing leaders to make more informed choices. By generating comprehensive reports that synthesize complex information into actionable insights, AI helps leaders navigate ambiguity and make decisions with greater confidence.

Innovation acceleration is another area where GenAI can have a significant impact. Leaders can use AI-generated ideas as a starting point for brainstorming and product development, expanding the realm of possible solutions. By automating routine tasks, AI frees up human creativity for higher-value innovation activities. The ability to rapidly prototype and test new ideas through AI-powered simulations can significantly speed up the innovation cycle, allowing companies to bring new products and services to market faster.

In the realm of customer experience, GenAI enables leaders to revolutionize how their companies interact with and serve customers. AI can create hyper-personalized marketing content and product recommendations, tailoring the customer experience to individual preferences and behaviors. More sophisticated and empathetic AI-powered customer service systems can handle complex queries, providing faster and more satisfactory resolutions. By anticipating customer needs through predictive analytics and generating proactive solutions, companies can stay ahead of customer expectations and build stronger, more loyal relationships.

Operational efficiency is another area where GenAI can drive significant improvements. In supply chain management, AI can optimize processes by predicting disruptions and suggesting alternatives, ensuring smoother operations. Complex processes, from financial forecasting to resource allocation, can be automated and optimized using AI, freeing up human resources for more strategic tasks. By generating optimized schedules and workflows, AI can improve overall productivity across the organization.

7. LEGAL CONSIDERATIONS FOR AI IN THE WORKPLACE

As organizations increasingly adopt AI and GenAI tools in the workplace, they must navigate a complex landscape of legal considerations. These technologies, while offering tremendous benefits, also present unique legal challenges that employers must address to mitigate risks and ensure compliance.

One of the primary legal concerns surrounds data privacy and protection. AI systems, particularly GenAI, require vast amounts of data to function effectively. Employers must ensure that their use of employee and customer data complies with relevant data protection laws, such as the General Data Protection Regulation (GDPR) in the E.U. or the California Consumer Privacy Act (CCPA) in the U.S. This includes obtaining proper consent for data collection and use, implementing robust data security measures, and providing transparency about how AI systems use personal data. These AI-related laws seem to be evolving almost as quickly as AI itself, so compliance is an ongoing requirement.

Employers must also be mindful of potential discrimination and bias issues when using AI in employment decisions. AI is trained using data that often contains human biases, so those biases are often present in AI output. Consequently, AI systems used for recruitment, performance evaluation, or promotion decisions could inadvertently perpetuate or even exacerbate existing biases if not carefully designed and monitored. In the U.S., for example, the use of AI in employment decisions

must comply with federal anti-discrimination laws such as Title VII of the Civil Rights Act, the Age Discrimination in Employment Act, and the Americans with Disabilities Act. There are also local laws such as the New York City Local Law 144 that regulate employers' use of augmented human resource related decision making. Employers should regularly audit their AI systems for potential bias and be prepared to demonstrate that their AI-driven decisions do not discriminate against protected classes.

Intellectual property rights present another significant legal consideration, particularly with the use of GenAI. When employees use AI tools to create content, questions may arise about who owns the resulting intellectual property. Employers should clearly define in their policies and employment agreements how AI-generated content will be treated in terms of ownership and usage rights. Additionally, organizations must ensure that their use of AI tools does not infringe on third-party intellectual property rights, as GenAI models may inadvertently reproduce copyrighted material.

The use of AI in workplace monitoring and surveillance also raises legal and ethical concerns. While AI can enhance productivity and security, excessive or covert monitoring may violate employees' privacy rights and damage trust. Employers must balance their legitimate business interests with employees' reasonable expectations of privacy. In many jurisdictions, employers are required to inform employees about the extent and nature of workplace monitoring and obtain consent where necessary.



Liability and accountability for AI-driven decisions is an evolving area of law that employers must closely monitor. As AI systems become more autonomous in decision making, questions arise about who is legally responsible when something goes wrong. For instance, if an AI system makes a decision that results in harm or loss, it may not always be clear whether the employer, the AI developer, or another party should be held liable. Employers should seek to clearly define lines of accountability and consider how their insurance policies cover AI-related risks.

The use of AI in certain regulated industries, such as healthcare or finance, may be subject to additional legal requirements. For example, in healthcare, AI systems that assist in diagnosis or treatment decisions may be considered medical devices and, therefore, subject to regulatory approval processes. In the financial services sector, AI systems used for trading or risk assessment may need to comply with specific regulatory standards for transparency and auditability.

As AI technology evolves rapidly, so does the legal landscape surrounding its use. Many jurisdictions are in the process of developing or updating laws and regulations specifically addressing AI. The proposed E.U. AI Act, for example, aims to create a comprehensive regulatory framework for AI systems based on their level of risk. Employers must stay informed about these evolving legal frameworks and be prepared to adapt their AI strategies accordingly.

7.1 Practical tips

To navigate these complex legal issues, organizations should consider the following steps:

1. Develop comprehensive AI governance policies that address data privacy, non-discrimination, intellectual property, and other relevant legal considerations.
2. Regularly conduct AI audits and impact assessments to identify and mitigate potential legal risks.
3. Provide training to employees on the legal and ethical use of AI tools in the workplace.
4. Engage legal experts specializing in AI and technology law to stay abreast of legal developments and ensure compliance.
5. Maintain open communication with employees about the use of AI in the workplace, addressing concerns and fostering trust.

By proactively addressing these legal considerations, employers can harness the benefits of AI and GenAI tools while minimizing legal risks and building trust with their workforce and stakeholders.

7.2 Other challenges and considerations

While the potential of GenAI is immense, its widespread adoption also brings significant challenges that leaders must address. Ethical concerns and governance issues become more pressing as AI systems become more advanced and autonomous. Questions about decision making transparency, potential biases in AI algorithms, and accountability for AI-driven decisions need careful consideration. Leaders must establish governance frameworks to ensure responsible AI use, balancing innovation with ethical considerations and societal impact.

The potential for job displacement and workforce transition is a significant concern. While new jobs will be created in the AI era, there is a risk of short-term displacement in certain sectors. Leaders must manage this transition sensitively, balancing efficiency gains with social responsibility. This might involve investing heavily in reskilling programs, creating new roles that leverage human-AI collaboration, and providing support for employees whose roles are significantly impacted by AI adoption.

Data privacy and security concerns are amplified in the age of GenAI, which requires vast amounts of data to function effectively. Leaders must ensure that stringent data governance practices are in place, protecting both customer and employee data. This involves not only complying with data protection regulations but also being transparent about data usage and implementing robust cybersecurity measures.

Quality control and reliability of AI-generated content and AI-driven actions is another crucial consideration. We have all heard about hallucination, where AI simply makes stuff up. Leaders must implement testing and validation processes to ensure the accuracy and reliability of AI outputs. This is particularly important in industries where AI decisions can have significant consequences, such as healthcare or finance.

As regulatory frameworks evolve to keep pace with AI advancements, staying compliant becomes increasingly complex. Leaders must stay informed about emerging AI regulations and ensure their AI initiatives comply with current and future legal requirements. This might involve working closely with legal teams, participating in industry discussions on AI governance, and advocating for balanced regulations that promote innovation while protecting societal interests.

8. CONCLUSION: EMBRACING THE GenAI FUTURE

The rise of GenAI marks a new chapter in technological advancement, promising to reshape not just how we work, but how we create, innovate, and solve complex problems. For business leaders, this presents an unprecedented opportunity to drive growth, innovation, and competitive advantage. However, success in this new era requires more than just technological adoption. It demands a fundamental shift in leadership approach – one that balances technological innovation with human-centric values, ethical considerations, and a commitment to continuous learning and adaptation.

Leaders who can effectively harness the power of GenAI while nurturing human creativity, empathy, and ethical judgment will be the architects of tomorrow's most successful and resilient organizations. They will create workplaces where humans and

AI systems collaborate seamlessly, each amplifying the other's strengths. As we navigate this transformative era, the role of leadership in guiding organizations through these changes cannot be overstated.

The GenAI revolution is not just about technology, it is about reimagining our relationship with work, with each other, and with the world around us. As we embrace this new era, we have the opportunity to shape a future where technology enhances human potential, creates new possibilities, and contributes to a more prosperous and equitable world. The journey of AI integration is just beginning, and the coming years will likely bring even more revolutionary advancements. In this rapidly changing landscape, agility, ethical consideration, and a commitment to continuous learning will be the cornerstones of success. For leaders willing to embrace this challenge, the GenAI era offers an exciting opportunity to make a lasting impact on their organizations and society at large.

REFERENCES

- Achenbach, J., K. Arbeiter, N. Mellors, and R. Shahani, 2024, "Harnessing Generative AI in manufacturing and supply chains," McKinsey & Co., <https://tinyurl.com/2acc4c7d>
- Alowais, S. A., et al., 2023, "Revolutionizing healthcare: the role of artificial intelligence in clinical practice," *BMC Medical Education* 23:689, <https://tinyurl.com/8cebwwc4>
- Burke, L., 2024, "Employers find openings to share AI bias liability with vendors," *Bloomberg Law News*, July 15, <https://tinyurl.com/4sx8xx26>
- Haythornthwaite, R., and N. Pierce, 2019, "Fourth revolution board of director imperatives: artificial intelligence, robots, reskilling and ethics," *QIO & Littler Mendelson*, May
- Kempe, L., 2024, "Navigating the AI employment bias maze: legal compliance guidelines and strategies," *ABA Business Law Section*, April 10, <https://tinyurl.com/yw6vfy5>
- Lawrence, A., 2024, "AI's impact on robots in manufacturing," *American Machinist*, September 10, <https://tinyurl.com/2s43n278>
- Marr, B., 2024, "How generative AI is accelerating drug discovery," *Forbes*, June 19, <https://tinyurl.com/nw8m2wh9>
- McKinsey & Co., 2023, "The economic potential of generative AI: the next productivity frontier," McKinsey Global Institute, June, <https://tinyurl.com/5vj3rvb>
- Ocampo, D., 2024, "CCPA and the E.U. AI Act," *California Lawyers Association*, June, <https://tinyurl.com/32nm6bh2>
- Payne, D., 2024, "Who pays when AI steers your doctor wrong?" *Politico*, March 24, <https://tinyurl.com/2p9kc9se>
- Pierce, N., and S. Goutos, 2023a, "AI at work: building a future-ready workforce," *Gunderson Dettmer*, December 13, <https://tinyurl.com/2s4arr8w>
- Pierce, N., and S. Goutos, 2023b, "Why lawyers must responsibly embrace generative AI," *Berkeley Business Law Journal* 21:2, <https://tinyurl.com/y37dc7wm>
- Rainie, L., M. Anderson, C. McClain, E. A. Vogels, and R. Gelles-Watnick, 2023, "Report: AI in hiring and evaluating workers: what Americans think," *Pew Research Center*, April 20, <https://tinyurl.com/4mm8skux>
- Son, H., 2023, "Morgan Stanley is testing an open AI-powered chatbot for its 16,000 financial advisors," *CNBC*, March 14, <https://tinyurl.com/4t77cfdp>
- Stone, M., 2023, "eBay's first Chief AI Officer is building AI tools to change how people shop online," *Business Insider India*, October 10, <https://tinyurl.com/3sdnepda>
- van den Berg, G., 2024, "Generative AI and educators: partnering in using open digital content for transforming education," *Open Praxis*, April 3, <https://tinyurl.com/3c96dpwk>
- Westfelt, A., and N. Pierce, 2023, "Client insight: legislating the future of AI in employment: NYC's law on automated decision tools and other important developments," *Gunderson Dettmer*, July 26, <https://tinyurl.com/4evcpnse>
- Yu, H., and Y. Guo, 2023, "Generative artificial intelligence empowers educational reform: current status, issues and prospects," *Frontiers in Education* 8, <https://tinyurl.com/yc74dx3z>



ORGANIZATION

- 56 How corporate boards must approach AI governance**
Arun Sundararajan, Harold Price Professor of Entrepreneurship and Director of the Fubon Center for Technology, Business, and Innovation, Stern School of Business, New York University
- 66 Transforming organizations through AI: Emerging strategies for navigating the future of business**
Feng Li, Associate Dean for Research and Innovation and Chair of Information Management, Bayes Business School (formerly Cass), City St George's, University of London
Harvey Lewis, Partner, Ernst & Young (EY), London
- 74 The challenges of AI and GenAI use in the public sector**
Albert Sanchez-Graells, Professor of Economic Law, University of Bristol Law School
- 78 AI safety and the value preservation imperative**
Sean Lyons, Author of Corporate Defense and the Value Preservation Imperative: Bulletproof Your Corporate Defense Program
- 92 Generative AI technology blueprint: Architecting the future of AI-infused solutions**
Charlotte Byrne, Managing Principal, Capco
Thomas Hill, Principal Consultant, Capco
- 96 Unlocking AI's potential through metacognition in decision making**
Sean McMinn, Director of Center for Educational Innovation, Hong Kong University of Science and Technology
Joon Nak Choi, Advisor to the MSc in Business Analytics and Adjunct Associate Professor, Hong Kong University of Science and Technology

HOW CORPORATE BOARDS MUST APPROACH AI GOVERNANCE

ARUN SUNDARARAJAN | Harold Price Professor of Entrepreneurship and Director of the Fubon Center for Technology, Business, and Innovation, Stern School of Business, New York University

ABSTRACT

As the landscape of artificial intelligence (AI) evolves rapidly, AI oversight by corporate boards is essential for managing AI exposure and complying with new AI laws. Competitive pressure to stay ahead in the AI race is inducing CEOs to embrace innovation aggressively, making board oversight especially critical. This paper presents a framework for corporate boards that identifies some key AI governance dimensions and provides guidelines for assessing their organizational risk and regulatory likelihood. The dual lenses of risk and regulation can simultaneously aid a board in prioritizing governance aspects to pay attention to and in choosing a robust oversight strategy. Mapping the risk-regulation matrix shapes appropriate recommended oversight strategies, ranging from proactive self-regulation and compliance monitoring to more passive wait-and-watch strategies. The paper further provides a structured way to navigate the evolving regulatory and governance landscape while unshackling boards from the subjectivity and imprecision of terms like “responsible” or “ethical” AI, leading to oversight that aligns with a company’s unique risk profile and industry-specific regulatory context, while recognizing that AI governance touches a range of topics, from technology, intellectual property and sustainability to audit, measurement, and risk assessment.

1. INTRODUCTION: THE EVOLVING LANDSCAPE OF AI GOVERNANCE

The landscape of AI governance has become decidedly more multifaceted over the last two years. Before 2022, two issues – data privacy and algorithmic bias – were a primary focus of both internal corporate governance and government legislation efforts. Most saliently, the E.U.’s General Data Protection Regulation (GDPR) redefined consumer data protections globally, while also introducing key E.U.-specific requirements on algorithmic profiling, the transparency of algorithms, and the detection of potential biases in automated decision systems. GDPR inspired parallel legislation in countries ranging from the U.K. (the 2018 Data Protection Act) to Brazil (the 2020 LGPD¹). In parallel, China’s 2021 Personal Information Protection Laws required that the use of personal information in automated decision making does not lead to discriminatory treatment. While the ambitious Algorithmic Accountability Act in the

U.S. is unlikely to become federal law, a growing number of state and local laws (including, for example, New York City’s Local Law 144 – 2021) are mandating actions to mitigate algorithmic bias. Meanwhile, long-standing anti-discrimination laws in many countries have translated into requirements that machine learning systems not use “protected attributes” as features of training data, and there have been self-regulatory efforts by organizations worldwide to minimize the replication of these attributes from combinations of other training data features.

Generative AI (GenAI) has made this governance landscape substantially more complex. The inherent unpredictability of GenAI creates an array of issues of robustness: occasional “hallucinations” in AI output are now a reality that must be managed rather than an error that can be corrected, and generated content must align with organizational brand. Blurring lines between the quality of human- and AI-generated

¹ Lei Geral de Proteção de Dados Pessoais

content raises the question of whether an organization must be transparent about attributing machine-generated content. More broadly, the notion that one can aspire to make one's AI "transparent" is an increasingly utopian ideal in an era of large language models (LLMs) with trillions of parameters. There are new governance issues around appropriate training data for LLMs, from copyright infringement to the leakage of corporate intellectual property. The enormous energy needs of AI infrastructure challenge sustainability goals, while workforce displacement issues seem poised to take center-stage as the capabilities of AI become more human-like. Meanwhile, the challenges of fairness and privacy remain: the ascendance of GenAI has raised novel and subtle possibilities for unintended bias, while discussions around data privacy have become more nuanced, with separate attention needed to consumer data protection, training data governance, and preserving the intellectual autonomy of human workers.

Many excellent and current AI governance guidelines exist for governments and policymakers.² However, for a corporate board, navigating oversight in this multifaceted and evolving governance environment is a significant challenge. Some boards struggle to assess whether AI governance is a strategic role or a control role, and whether a dedicated new AI committee is necessary or if AI-related oversight can be subsumed by standing risk or audit committees. Broad subjective phrases like "responsible AI" and "ethical AI" induce lengthy discussions about the scope of what constitutes responsible or ethical behavior while compounding uncertainty about the connection of responsible AI to broader corporate social responsibility.

As this article will explain, breaking down AI governance into its specific dimensions can significantly enhance clarity, and assessing each of these dimensions through the dual lenses of risk and regulation can simultaneously aid a board in prioritizing them and in choosing a robust oversight strategy.

2. SOME KEY DIMENSIONS OF AI GOVERNANCE

The set of specific issues that might fall under the broad umbrella of AI governance is evolving. I discuss some of today's most salient dimensions in what follows. These are arranged in no particular order, and as I will explain later, there is no ranking of importance inherent in the order in which they are presented. Put differently, there is no absolute prioritization

– relative importance is specific to an organization, and further, can emerge only from a process of assessing risk, reinterpreting existing laws in the AI context, and anticipating industry-specific regulation.

2.1 AI alignment

The use of AI implies a ceding, to varying extents, of autonomy in what the humans in an organization do. This makes it important to ensure that this autonomy does not lead to a divergence between organizational values, goals, or culture and the choices made by AI systems. A useful dichotomy is between "content alignment" and "decision alignment".

- **Content alignment:** involves ensuring that the generated "content" of an AI system is aligned with an organization's objectives or principles. For companies like Google or OpenAI that create general-purpose GenAI, this involves ensuring that AI output does not inadvertently create unacceptable content ranging from hate speech to prohibited topics. For most other companies that adapt these GenAI systems into business applications, content alignment will focus more on ensuring that the output of these applications, whether from a conversational AI system interacting with clients or a system being used to generate marketing content, is aligned with the brand and image of the organization.
- **Decision alignment:** involves ensuring that "decisions" that are delegated to an AI system are aligned with organizational goals. Such alignment has for many years been the focus of companies creating self-driving automobiles and have brought philosophical discussions like those of the "trolley problem" into mainstream business debates.³ For most other companies, issues of decision alignment may be more frequent in lower stakes situations – for example, about the nature of decisions a customer service chatbot makes about product refunds or rebates when conversing with a customer, or decisions about recommendation/advertising targeting.

2.2 Intellectual property (IP) governance

To understand the most important IP governance issues related to AI, one must first recognize that the growing scale of AI systems leads to a build-versus-buy managerial assessment that is elevated to being a governance issue due to the proliferation of open-source models like the Llama LLM series released to the public by Meta (formerly Facebook) and a range

² <https://tinyurl.com/mrke9tu>

³ <https://tinyurl.com/y965aen2>

of models developed by academics and others available on repositories like Hugging Face. Choosing open source is cost-effective, allows greater IP control over customized systems, and places transparency choices more squarely in the hands of the organization. However, it can create quality control and security issues,⁴ and can require in-house AI talent beyond the reach of many, impeding future progress for an organization not on the scientific cutting-edge of AI. In contrast, relying on a vendor like OpenAI or Google can be extremely expensive as an organization's AI use scales, can lead to opaqueness being a default rather than a choice, and, in some cases, may lead to lock-in that can constrain innovation and increase future cost uncertainty.

A deeper IP issue arises when one unpacks how shared GenAI technologies play a growing role in building AI applications for specific organizational uses. We are accustomed to AI systems being trained on structured sets of proprietary outcomes. However, large language models (LLMs) and other GenAI systems for images and video are trained on massive datasets that often encompass the entirety of humanity's available digitized content. For example, it is believed that OpenAI's GPT models are trained on all publicly available digital written content. Now consider the typical way in which most organizations will adapt a general-purpose system like LLMs for their specific purposes (for example, to create a customer service chatbot that understands the company's products, or an AI system for employees knowledgeable about the company's human resources policies and practices). One approach involves customizing an LLM developed by a company like OpenAI or Google using corporate specific knowledge (a process called "fine-tuning"). Although the AI systems that emerge from this process are proprietary to the company, corporate IP has, in a sense, been absorbed into the model's parameters. A different approach involves "augmenting" what is sent to a (non-proprietary) LLM with fragments of internal documents or past relevant conversations "retrieved" from an internal knowledge management system (a design often implemented using what is called "retrieval augmented generation" or RAG). Again, unless the company develops and hosts its own LLM, company knowhow is being sent (albeit in small chunks) to an external system. Whichever strategy a company chooses, the IP challenge is clear – this kind of tacit knowledge transfer requires careful oversight and thought.

2.3 Training data governance

The governance issues around training data that lead to the creation or use of a company's AI systems do not stop with the IP challenges discussed above. Oversight of the possible liabilities that a company may face on account of the training data used in its AI systems is also essential. Again, this is a multifaceted issue.

- An organization must determine the extent to which it is aware of all the data that may have been used to train the systems used by its AI applications. If using shared GenAI infrastructure like OpenAI's GPT or Google's Gemini, it must also consider whether to be prepared for regulatory demands that associated training data be made "transparent", either to a regulator or to the public.
- It is also almost certain that the training datasets of all LLMs and image generating systems have included "copyrighted" information used without the explicit permission of the copyright holders.⁵ Although courts in the U.S. may eventually deem this use of copyrighted content "fair use", this is neither certain nor internationally applicable. Some countries like Singapore already have explicitly legislated the use of copyrighted information for AI model training, others like Australia have far more restrictive definitions of fair use than that of the U.S. The uncertainty and variance in how different countries will resolve the question of fair use makes this a key governance issue, since the direct liability associated with regulatory shocks could be significant. Even if an organization is not training its own LLMs, there may be substantial indirect costs if these shocks lead to unexpected changes in the availability or performance of the LLMs that one's AI systems depend on.
- Over time, organizations will increasingly use the output of their employees as training data for new or improved AI systems. For example, employees may be permitted to create "digital twins" that raise productivity by writing in their style or voice, draft contracts, or serve as chatbot substitutes when the employee is unavailable. Although this idea of a digital replica may seem like science fiction, it is increasingly feasible with today's AI technologies. An organization that is capturing and encoding the human capital of its workforce in AI systems must think through and implement a framework that regulates use, longevity, and value sharing from such systems.

⁴ <https://tinyurl.com/4drcjxjb>

⁵ <https://tinyurl.com/yu7kadd9>

2.4 Model explainability

Boards must often contend with the extent to which they insist that the AI systems their organization uses generate output whose logic can be explained. Over the last 20 years, artificial multi-layered neural networks (often called “deep learning” systems)⁶ have become the favored model for building machine learning systems. Their superior performance comes with a hidden cost, because “explaining” the logic of their underlying statistical models is impossible. For example, an organization using a deep learning system for loan approval may be unable to explain why the system turned down a specific loan application. In contrast, a simpler underlying model based on logistic regression⁷ that places weights on different features could allow an organization to explain that it was the income level or the credit score that led to the decision, but such an explainable system will almost certainly make less profitable decisions. This landscape is further complicated by GenAI systems, not just because their workings are not amenable to explainability, but because it is highly likely than any organization that is not a tech giant is instead reliant on systems built by companies such as OpenAI, Google, Anthropic or Meta, and is thus limited in its quest for explainability by the choices made by its AI vendors.

2.5 Model transparency

Independent of explainability, an organization may face internal or external pressure to make the workings of its AI systems “transparent”. For example, in its early days, Uber faced pressure to make the details of its surge pricing algorithm visible to users and regulators. An insurance company using AI to price its products and set premiums may consider whether to explain the logic of this process to all its consumers. Similarly, an investment firm using AI to make trading decisions may face transparency pressure from regulators towards creating a system-wide view to assess contagion risks. Beyond the explainability-performance trade-off associated with neural networks, transparency can have competitive impacts as the performance of the AI systems becomes an increasingly important source of advantage. And again, the transparency options of an organization will be limited by the transparency choices made by its GenAI vendors like OpenAI, Google, Anthropic, or Meta.

2.6 AI robustness

AI has always been less predictable than its deterministically programmed counterparts. This is a natural consequence of the paradigm – a machine learning system that makes predictions based on a probabilistic statistical model will always have some associated unpredictability. There is no absolute way around this trade-off because a completely predictable machine learning system has little value – the unexpectedness of predictions and their departures from what human analysts may come up with is what makes them useful.

As the underlying statistical models have become larger and more complex, the associated unpredictability has grown. It is widely recognized that LLMs tend to “hallucinate”, confidently providing information that is imagined and incorrect. Since LLMs generate new and original content through a process of successive next-word prediction,⁸ these hallucinations will never be eliminated and must instead be managed. The governance of AI robustness thus involves balancing the trade-off between creativity and human-likeness on the one hand, and accuracy on the other, especially for AI systems that are customer-facing. Appropriate systems for conflict resolution and due process must be determined if, for example, a customer is provided with incorrect information about a refund or an interest rate by a customer service chatbot, or an employee makes vacation plans based on an outdated policy provided by an internal LLM-based human resources system.

A related dimension of robustness will relate to managing more subtle “traits” of underlying LLMs, especially in an environment where new versions are released with increasing frequency. These new versions typically report improved performance based on a variety of standardized benchmarks. Applications built on top of LLMs must then decide whether to take advantage of these improvements or stay with a tried-and-tested older model, a decision often taken without clarity about more subtle behavior changes that the transition may induce. Recent research⁹ has shown, for example, that the Fall 2024 version of OpenAI’s GPT4 (named o1), while outperforming its predecessor on most standardized metrics, demonstrates a dramatic drop in the human-like trusting behaviors that it displays. As LLMs form the basis for a growing number of high-stakes commercial systems, their increasing opacity and complexity can lead to hidden fault lines, adding another layer of complexity to the governance of AI robustness.

⁶ <https://tinyurl.com/mrxamrj7>

⁷ <https://tinyurl.com/3k5vke35>

⁸ <https://tinyurl.com/mvyejty6>

⁹ Li, Sedoc, and Sundararajan, unpublished.

2.7 Machine attribution

One of the most common uses of GenAI is to generate new written and visual content, from marketing and advertising material to customer communications. Video generating AI will soon be ubiquitous. Large-language models also excel at summarizing written content. Granted, tactical decisions about the right mix of human- and machine-generated materials may receive executive focus organically, but there is an associated governance choice of attribution – whether to reveal the AI versus human provenance of public-facing content, and if so, in what situations. It may seem natural to label an AI-generated video as having been AI-generated, but what about an AI-generated summary of user reviews, a marketing document that was generated with the aid of AI but with some human participation, or an advertising image that was human-created but hyper-personalized using AI? Insufficient machine attribution could lead to customer backlash, while excessive attribution could create the impression of inauthenticity.

Additionally, as AI agents take over larger fractions of synchronous and conversational customer interaction, a related machine attribution issue that requires clear governance is whether to inform a customer when they are interacting with an AI agent rather than a human. Today, most AI-driven customer interaction systems, from automated voice systems to website chatbots are easily recognizable as being non-human. As the human-machine lines blur in the coming years, many of these choices will be driven by government regulation, since this is an issue high on the legislative agenda, but boards must nevertheless proactively ensure that their organization makes choices on this front that are aligned with their brand and values.

2.8 Algorithmic bias and inclusivity

AI systems tend to reflect, or even amplify, the biases present in the data they are trained on. In simple terms, absent active intervention, biases that exist in society – whether related to gender, race, or socioeconomic status – can be inadvertently encoded into AI systems. For example, an AI-driven recruitment tool might favor candidates of a certain background because it was trained on historical hiring data that reflected existing inequalities. This issue has grown in prominence as AI has taken on greater decision making roles in areas like hiring, lending, and law enforcement.

Bias in AI systems is not a new issue. As machine learning proliferated in real-world settings, the potential to reproduce discriminatory outcomes has been widely recognized over the past decade. A variety of cases have received extensive media coverage, from predictive policing tools unfairly targeting certain communities and bail decision systems possibly displaying bias in denial to healthcare algorithms exhibiting racial biases in treatment recommendations.

With the emergence of GenAI, however, these challenges have taken on new dimensions. As discussed, LLMs create new content after being trained on large, diverse datasets. Their training data includes vast amounts of internet data, unmoderated content with a higher likelihood of biased views and dialog. Thus, the parameters in a GenAI model might reflect cultural stereotypes and gender biases that are subtle but eventually have widespread influences. It is difficult to isolate and address these biases by altering training datasets due to their enormity and opaqueness.

While a board might simply be tempted to ask that their GenAI be created in a way that aligns its “views” with the organization’s values, LLMs operate in a way that makes it difficult to directly change their “beliefs”. Unlike a human being, an LLM does not consciously hold beliefs; instead, it generates responses based on statistical associations derived from training data. As a result, when a generative model starts displaying biased behavior, there is no direct way to correct its underlying worldview. Instead, developers are forced to add increasingly complex sets of guardrails – specialized programs and machine learning systems that check output – to try to mitigate harmful or biased outputs. These guardrails involve varied techniques and policies that attempt to filter or guide the responses generated by the model. While these methods can be effective to some extent, they are not foolproof, and surrounding an AI system with an increasingly complex web of guardrails increases its fragility.

2.9 AI use and sustainability

AI consumes a growing fraction of the electricity of countries in which its hardware infrastructure is based. By some estimates, the power needs of AI in the U.S. will eventually exceed those of New York City, and AI data centers are projected to constitute close to 40% of the total increase in U.S. power demand by 2030.¹⁰ For AI producers like OpenAI, Microsoft, Google, and Meta, this already creates a significant

¹⁰ <https://tinyurl.com/5dp39r5z>

sustainability challenge. For example, since ChatGPT was released, Microsoft has scaled back and fallen short of its sustainability goals,¹¹ while aggressively seeking out alternative sources of sustainable power, including recently striking a deal to use the entire 837MW output of the fabled and recently recommissioned nuclear power plant at Three Mile Island in Pennsylvania.¹² Every organization must assume that their AI usage will grow dramatically in the coming years, and that each new generation of AI will be increasingly power-hungry. Examining the sustainability footprint of one's AI providers while balancing the quest for innovation with the organization's sustainability goals requires careful thought and oversight.

2.10 AI workforce displacement and transition planning

It is widely anticipated that changes in the mix of activities between machines and humans will cause a significant percentage of the workforce in the U.S., Western Europe, and Japan to transition to a new occupation over the coming two decades. Some estimates suggest that by 2030, one in 16 workers will need a new occupation due to AI workforce displacement.¹³ Corporations must decide how proactive to be in supporting their employees to adapt, grow, and invest in their skills.

"Reskilling" is something that is seen as a cost driver today but may be central to a brand's identity in the future. A useful parallel comes from how corporate approaches to sustainability or responsible labor practices have evolved. A couple of decades ago, both were seen as part of corporate social responsibility, choices that drove up costs rather than profits. Today, people make consumption choices based on a brand's sustainability positioning and may shun companies with unfair labor practices. A decade from now, the same may be true about responsible workforce transition policies.

Educational funding from governments has traditionally focused on early-career development. One might argue that corporations are uniquely positioned to create opportunities for mid-career reskilling that align directly with their evolving needs. However, this requires more than just offering skill-based training programs. Just as universities prepare students for their first careers with a broad range of experiences beyond the classroom, corporations should build reskilling programs

that go beyond mere technical training. These programs should include mentoring, career coaching, networking opportunities, and branded credentials. By providing these additional elements, corporations can help employees build confidence, develop professional networks, and explore new career paths.

A deeper governance issue relates to human intellectual autonomy.¹⁴ Today's AI technologies hold the potential to decentralize access to a wide range of skills and productive capabilities, empowering millions to follow entrepreneurial pursuits while fostering the rise of a new generation of AI-driven professionals – from educators and healthcare providers to investment advisors and data scientists. As discussed briefly in Section 2.2, as AI systems within an organization progressively encapsulate the human capital of a workforce, if individuals cannot assert a level of ownership over their personal generative processes, talents, or expertise, we may face a future where intelligence and skills become overly commoditized and centralized. This could render humans unable to reap the economic rewards of their own human capital investments, reducing the benefits of AI to a select few rather than the broader population.

While this list of governance issues is lengthy, it is by no means exhaustive. For example, a board must consider how AI changes its existing governance approaches to cybersecurity and data privacy. And over time, new AI capabilities are bound to bring new governance challenges. Addressing them requires a nuanced assessment of organizational risk and a delicate balance between self-regulation and compliance. I unpack these points in greater detail in the following section.

3. AI GOVERNANCE OVERSIGHT: RISK AND REGULATION

Oversight of all these dimensions of AI governance can be a challenge for any board. To prioritize, each AI governance dimension should be evaluated through two critical lenses.

The first lens is the level of risk that the AI governance issues associated with a dimension might pose to one's specific organization. For example, there may be little or no actual risk posed to an organization that does not operate in the technology space if they choose not to make transparent the fact that they are using publicly available training datasets. In

¹¹ <https://tinyurl.com/ym6zpm66>

¹² <https://tinyurl.com/3fu2674z>

¹³ <https://tinyurl.com/mr2h7pvd>

¹⁴ <https://tinyurl.com/mtufyu5v>

contrast, leaking of key proprietary intellectual property due to flawed choices associated with letting a vendor fine-tune a version of their LLM to create a customer service chatbot could be quite damaging. Clearly, for dimensions that pose a higher organizational risk, careful thought must be given to risk mitigation strategies and a higher level of oversight is warranted.

The second lens is the likelihood that the dimension will be subject to government regulation in the near future. For example, it is highly likely that there will be government regulation relating to machine attribution from several agencies and jurisdictions. In contrast, it is unlikely that governments will find it necessary to create new legislation relating to the boundaries around a company's IP ownership when their data is used to train an AI system, tending instead to rely initially on existing IP laws and the bilateral contracting regime.

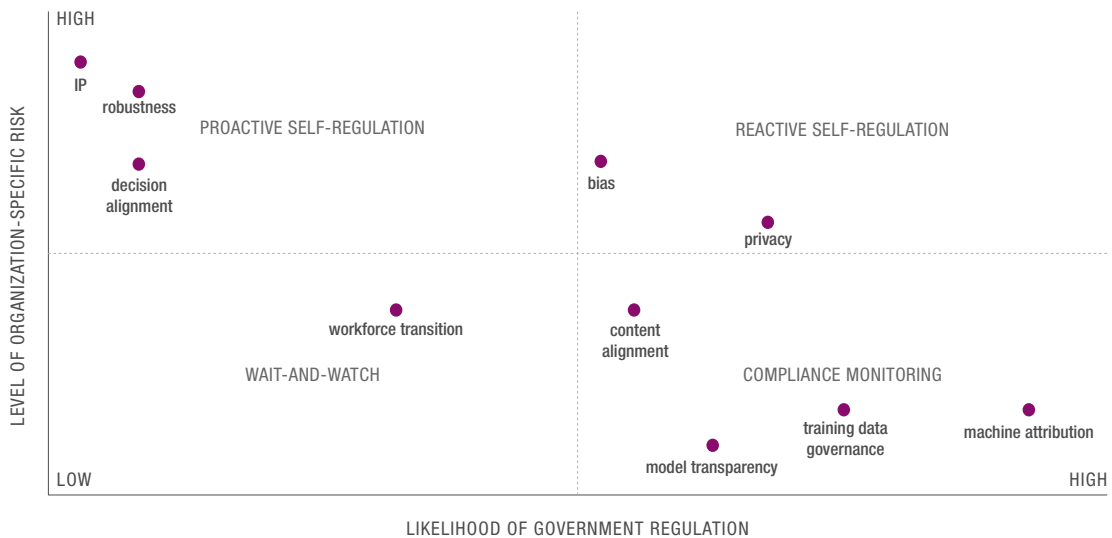
Placing each governance dimension according to its relative risk and regulatory likelihood clarifies the landscape of AI governance for a corporate board. An illustrative example of such a mapping is provided in Figure 1.

Importantly, there is no universal placement of these governance dimensions in the risk-regulation space – this is necessarily an **organization-specific** assessment. For example, an AI vendor like OpenAI faces significant risks associated with training data governance, while a company

in the oil industry may face lower risks on this dimension. Similarly, it is unlikely that governments will demand model transparency from the customer service chatbots of a consumer packaged goods company, but more likely they will consider this for AI systems that interact with financial markets and whose actions may affect the risk of financial contagion.

For a technology giant like Google or Meta, content alignment represents a high-risk dimension because the company's AI-generated content is widely disseminated and has the potential to have significant repercussions if untrue or misaligned with a country's value system. In contrast, a financial institution like a bank may view decision alignment as a higher-risk dimension because the decisions made by AI systems in the context of lending, risk assessment, or customer service can have direct and profound impacts on customers, regulatory compliance, and financial stability. Similarly, model explainability might be a relatively low-risk dimension for a manufacturing organization that uses AI primarily for internal process optimization. However, for an insurance company using AI to set premiums or approve claims, model explainability could be crucial, given the need to explain decisions to both customers and regulators. Similarly, AI robustness may be a top priority for companies developing mission-critical AI systems, such as those in aerospace or autonomous driving, where failure could have catastrophic consequences, while those in industries like retail, where AI use is largely for product recommendations

Figure 1: The risk-regulation matrix for AI governance



and targeting advertisements, this dimension might be important but not existential, allowing for a more measured approach to governance.

Depending on where each dimension lands, there are four broad oversight strategies that a board can consider.

3.1 Wait-and-watch

If a governance dimension is assessed as having both **low** organizational risk and a **low** likelihood of regulation, the recommended approach is to “wait and watch”. In this scenario, boards should do some planning and monitoring but prioritize the dimension lower on their governance agenda. For instance, consider the dimension of AI and sustainability for an organization whose AI use is not especially resource-intensive. Choices relating to the source of electricity used in this company pose low risk, and it is unlikely that there will be new pertinent regulation targeted specifically at the sustainability of the power used specifically for AI. The “wait and watch” approach allows a board to focus its governance attention elsewhere while staying informed about potential future shifts.

Of course, adopting a “wait and watch” strategy does not mean neglecting the governance dimension entirely. Monitoring the pulse of technological advancements that might affect the dimension is important. For example, five years ago, AI robustness was not on the radar of most companies or governments, but the recent rapid advances in GenAI have moved it on to the front burner.

3.2 Compliance monitoring

When a dimension presents **low** organizational **risk** but carries a **high** likelihood of **regulation**, boards should adopt a “compliance monitoring” approach. The goal here is to anticipate regulatory requirements and ensure the organization is ready to comply once those requirements are formalized. Boards might also consider whether compliance is likely to involve sufficiently high costs to warrant participating in the shaping of eventual regulation.

Machine attribution serves as a good example of a dimension in this category for many organizations, wherein absent regulation, the risks associated with attributing content as AI-generated, rather than human-created, are relatively low, especially if the content is non-sensitive or non-public-facing. However, driven by concerns about transparency and misinformation, governments are gradually requiring the attribution of machine-generated content, perhaps viewing

it as low-hanging fruit and a relatively non-controversial way to dip their toes into AI regulation. As AI agents assume larger fractions of conversational customer interactions and are imbued with greater economic autonomy, machine attribution will remain high on the regulatory priority list. Thus, monitoring regulatory developments closely and establishing internal processes that can be scaled up for compliance is prudent. This might include tracking proposed regulations in key markets and maintaining flexibility in labeling content as AI-generated. The emphasis here is on efficient allocation of resources – preparing to comply without overcommitting to a dimension that presents limited internal risk.

3.3 Proactive self-regulation

For governance dimensions with **high** organizational **risk** but a **low** likelihood of **regulatory** intervention, boards must insist that their company be proactive about crafting an internal self-regulatory regime. Waiting for regulations that may never arrive or viewing these governance dimensions as lower priority because of their absence on the government regulatory radar would be a mistake. Instead, organizations must take the lead in assessing risks and defining a governance framework.

Decision alignment and intellectual property governance are prime examples of dimensions that fall into this quadrant for many companies. In sectors like finance or healthcare, decisions made by AI systems can have profound impacts on customers. The organizational risk associated with misaligned decisions is significant. Proactive self-regulation in this context involves active red-teaming to ensure that decision making by any new AI system is aligned with the organization’s values and strategic goals. Creating internal standards for decision making transparency, establishing protocols for human oversight, and implementing safeguards to ensure that AI decisions can be adjusted when necessary are additional tactical steps that could help.

In certain settings, a board might consider asking the organization to take the lead in setting self-regulatory standards for their industry or sector, and creating a coalition that shares the same self-regulatory approach. For example, a group of companies may have greater leverage than any individual one if, perhaps through an industry consortium, they define and dictate shared standards around the boundaries of corporate IP when models are fine-tuned or sensitive information is sent to an external LLM in a RAG implementation.

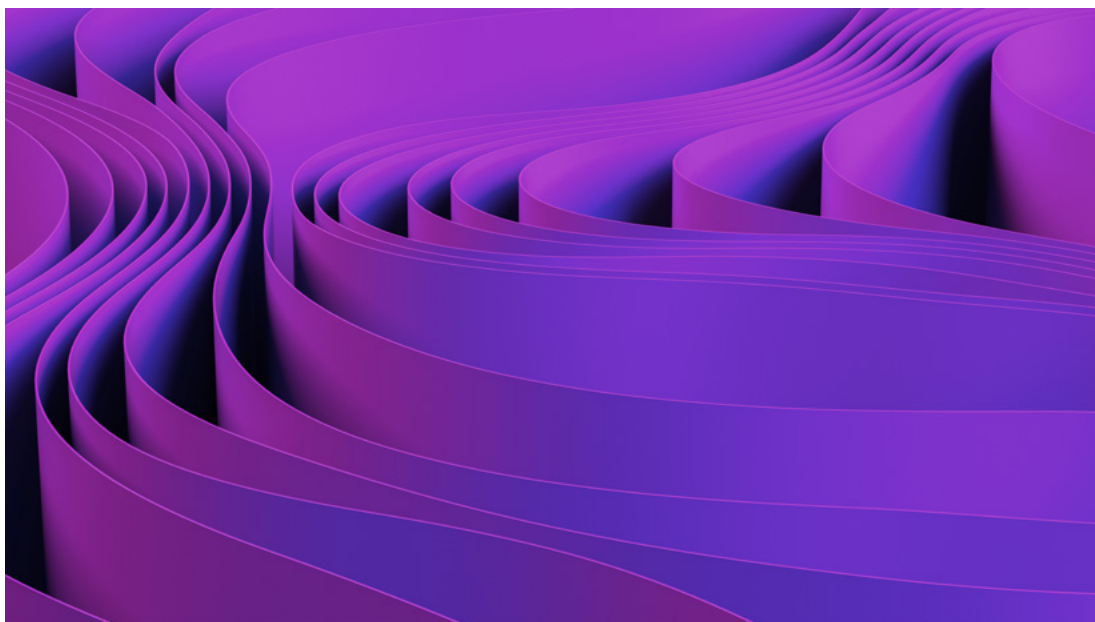
3.4 Reactive self-regulation

Finally, dimensions that exhibit both **high** organizational **risk** and a **high** likelihood of **regulation** must, of course, be given clear oversight priority, but a question that may arise is how to balance developing governance internally with anticipated external compliance. One approach would be to catalyze the active developing and implementation of internal governance policies, but to take a flexible rather than rigid approach while committing significant resources to shaping the expected government regulation towards aligning it with internal approaches. For a company like Google or Meta, who produce AI-generated content that reaches billions of users, ensuring that content is brand-aligned and does not inadvertently promote harmful or inappropriate material is both a high-risk issue and one already facing regulatory headwinds. In this case, internal steps like investing in content moderation technologies and establishing clear policies on acceptable content should be taken in parallel with active engagement with regulatory bodies to shape emerging standards. For governance dimensions in this quadrant, ensuring that internal self-regulatory approaches can be modified to meet new legal requirements as they emerge is crucial. Actively seeking dialogue with policymakers and contributing to, or leading, industry standards can also help align future regulations with existing internal practices, reducing the compliance burden associated with regulatory changes if they occur.

4. CONCLUSION: NAVIGATING AI OVERSIGHT

The framework provided in this paper lowers the complexity and obtuseness of AI governance by breaking it down into specific dimensions, an important first step towards prioritizing oversight. The dual lenses of risk and regulation can simultaneously help a board rank which aspects to pay attention to and choose a robust oversight strategy – from wait-and-watch and compliance monitoring for dimensions identified as having lower organizational risk to either reactive or proactive self-regulation for higher-risk dimensions, depending on the likelihood and imminence of government intervention. A board that merely monitors or discusses the latest AI legislation like the E.U.'s AI Act at a high level is providing insufficient oversight and control. Further, the relative prioritization of these different facets of AI governance must be specific to the company and industry. The importance of a tailored approach becomes apparent when considering that each organization has unique needs, hazards, and regulatory exposures, making it essential for boards to evaluate their specific context carefully.

Boards must aim to have at least one member sufficiently well-versed in the digital realm who can monitor the landscape and surface possible issues independent of the executive team. In parallel, the conversation about creating an AI governance committee should happen sooner rather than later. Many



organizations may be tempted to subsume AI oversight into an existing or standing committee like the audit committee, the risk committee, or the technology committee. However, as this article makes clear, AI governance overlaps with numerous specialized areas, from technology, intellectual property and sustainability to audit, measurement, and risk assessment. Having a dedicated committee lowers the risk of pursuing governance that is too deep and narrow, creates more robust oversight, and may be especially helpful in organizations with substantial AI investments or those operating in highly regulated industries. Such a committee can ensure that eventual AI governance has appropriately informed focus and control. A board-level committee can also facilitate a deeper understanding of emerging AI issues while ensuring that governance is balanced appropriately and judiciously with the executive team's desire to pursue more rapid or aggressive AI innovation.

Finally, boards would be well served by investing considerable thought during the phase in which they map their AI governance dimensions into the risk-regulation matrix, actively seeking appropriate executive, expert, and legal input to aid risk assessment and understand the likely legislative landscape. Elevating the importance of this step lends credibility to the idea that AI is a board priority and allows what follows to be undertaken with greater confidence. The ensuing oversight will then be targeted more prudently and boards can guide their management teams towards focusing executive attention where it matters most.

TRANSFORMING ORGANIZATIONS THROUGH AI: EMERGING STRATEGIES FOR NAVIGATING THE FUTURE OF BUSINESS

FENG LI | Associate Dean for Research and Innovation and Chair of Information Management,
Bayes Business School (formerly Cass), City St George's, University of London

HARVEY LEWIS | Partner, Ernst & Young (EY), London¹

ABSTRACT

The rise of artificial intelligence (AI), particularly generative AI (GenAI), presents both significant opportunities and challenges for business leaders. This paper explores how AI can reshape business models, operations, and the nature of work, drawing lessons from past technological revolutions and emerging insights from leading global organizations. It argues that AI's true potential lies not just in automating tasks but in fundamentally rethinking organizational processes and business models. The paper offers practical strategies for senior leaders to navigate this evolving landscape and successfully steer their organizations through an AI-driven future.

1. INTRODUCTION

There is so much conflicting information about the business implications of AI, particularly around generative AI (GenAI). On the one hand, investor enthusiasm and media hype remain high, fueled by numerous analyses projecting significant productivity gains.² On the other hand, concerns are mounting about the imminent bursting of the "AI bubble" and the subsequent "AI winter".

This stark contrast between hype and pessimism, with seemingly no middle ground, only adds to the confusion. This is not helped by the conflicting messages around "productivity" versus "transformation". Should AI be used simply to make existing work faster, or can it drive much deeper change within an organization, potentially altering business models, organizational processes, and the nature of work itself?

With such mixed signals, what strategies should senior business leaders adopt to navigate the future of business?

Drawing insights from how past technologies transformed businesses along with our ongoing research with leading organizations from the U.S., China, and Europe, we explore how AI transforms organizations and present new strategies for succeeding in an AI-driven future.

2. THE HYPE AND DESPAIR

There is no shortage of strong views about the future of AI. Leading consulting firms and investment banks, from McKinsey, BCG, to Goldman Sachs and JP Morgan, have made bold projections on multi-trillion-dollar additional economic growth from AI deployment, ranging from U.S.\$2.6 to U.S.\$4.4 trillion annually between now and 2030,³ to up to

¹ The views reflected in this article are the views of the author and do not necessarily reflect the views of the global EY organization or its member firms.

² While numerous analyses from institutions like the World Economic Forum and International Monetary Fund estimate AI's potential impact on productivity, these are often based on subjective judgments about AI's task capabilities. They also tend to overlook the fact that AI saving time on tasks does not automatically lead to increased productivity, as the time saved may not always be used productively.

³ <https://tinyurl.com/yxcph2vd>

U.S.\$16 trillion by 2030.⁴ Meanwhile, strong warnings have been made about the threats of AI, from privacy and security erosion, mass job displacement and economic disruption, to the ultimate threat to human existence.

According to Goldman Sachs, AI could replace the equivalent of 300 million full-time jobs globally by 2030.⁵ Discussions about universal basic income (UBI) also resurfaced, and Elon Musk was not alone in believing that due to the development of AI, “no job is needed”.⁶ The godfather of deep-learning, Geoffrey Hinton, went so far as to resigning from Google in order to speak freely about the danger of AI: “If I were advising governments, I would say that there’s a 10% chance these things will wipe out humanity in the next 20 years. I think that would be a reasonable number”. He went on to say that “[b]etween 5 and 20 years from now there’s a probability of about a half that we’ll have to confront the problem of [AI] trying to take over.”⁷

In practice, however, such warnings have largely been overshadowed by AI’s huge potential in improving productivity and transforming business and society. As Sam Altman jokingly remarked: “AI will most likely lead to the end of the world, but in the meantime, there will be great companies...”⁸

It is beyond the scope of this paper to adjudicate such debates. Our focus is on how AI transforms organizations and its strategic implications. We believe that the excessive focus on how AI will automate tasks and replace jobs is misguided. Indeed, simply automating tasks and jobs will not suffice to unlock AI’s full potential or justify the huge investments in its development and deployment. Moreover, current discussions of productivity enhancement via AI are not the same as “transforming business and society”.⁹ True transformation demands a more profound reimagining of organizational processes, business models, and the integration of human talent in an AI-driven world.

If history serves as our greatest teacher, then the true potential of AI will likely be realized through its ability to transform organizations and institutions – a process that historically takes decades rather than years. Meanwhile, it is important to note that while the transformation driven by AI may resemble past transformations, there are also reasons why it might be

different – faster, more uneven, and less predictable. These differences may arise because we are fundamentally dealing with a digital transformation rather than the mechanical transformations of the past. This has significant implications for the strategic approaches that senior business leaders should adopt to navigate the future of business

3. WHAT IS AI?

AI is commonly defined as technologies that mimic human intelligence and problem-solving abilities, but in practice, it often means different things to different people. The large variety of interpretations often lead to confusion in both casual conversations and formal decision making.

Today, much of the excitement about AI is related to AGI, or artificial general intelligence. Such AI will possess capabilities that are comparable or superior to humans in reasoning, conceptual learning, common sense, planning, cross-domain thinking, creativity, self-awareness, and emotions. However, AGI still does not exist today and there is no consensus on when (or if) it can be realized.

Today’s AI is far from AGI, even though it is important to note that in narrow slices of intelligence, it is already demonstrably superior to human capabilities. While this is not artificial “general” intelligence, we do not need it to be general before it is transformative. Narrow AI is already widely used in organizations, powering Google search, Amazon and Alibaba recommendations, and Uber and Didi ride-hailing matches, delivering significant efficiency gains and economic and social impact behind the scenes.

There are several popular AI classifications, but from a business perspective, it is useful to categorize AI in the current era into traditional analytical AI and GenAI. While analytical AI is widely used, GenAI is only beginning to make significant inroads in certain sectors (software development, media, creative services) and use cases (assistants/chatbots, for example).

Analytical AI, often referred to as “discriminative AI”, is primarily designed to follow predetermined rules and logic, and is widely used in applications that require consistency, precision,

⁴ <https://tinyurl.com/yvfz4e9j>

⁵ <https://tinyurl.com/4c6ru66z>

⁶ <https://tinyurl.com/4yfefys6>

⁷ Geoffrey Hinton during the Annual Romanes lecture at the University of Oxford in 2024: <https://tinyurl.com/4rjep5nz>

⁸ <https://tinyurl.com/ycx8pdpf>

⁹ When discussing productivity, the focus often shifts to AI’s potential to augment rather than replace workers. Some argue that we should be more candid about AI’s ability to replace people and take over many existing tasks, while also emphasizing its potential to create new work opportunities through AI-driven transformation.

and the ability to perform well-defined tasks. Examples include data analysis, decision making based on structured data, and rule-based problem solving. Analytical AI is often described as a fantastic “left brain”, as they are logical and precise.

In contrast, GenAI can generate new content or data that may or may not resemble the training data. Based on large language models (LLMs) powered by various architectures, including transformers, GenAI relies on machine learning models (particularly deep-learning and neural networks) to learn from vast amounts of data and generate outputs that are novel and original. GenAI is characterized by its adaptability, creativity, and capacity to handle ambiguous or incomplete information. Compared to analytical AI, GenAI is more effective for tasks requiring innovation and the generation of new ideas. Examples include applications that create text, images, music, and other forms of media. It has often been described as a new “right brain”, capable of creativity and generating new ideas.

It is important to note that despite their apparent differences, the distinction between these two forms of AI is becoming less clear, as LLMs are capable of either being analytical or of autonomously creating applications that are analytical.

The evolution of AI from systems that strictly adhere to human-defined rules to those capable of learning from data and creating new outputs is sparking significant new excitement. Analytical AI is foundational and remains crucial for large-scale applications in real operations, with most current AI applications relying on these technologies. GenAI is often viewed as a significant leap forward in AI’s ability to mimic human creativity and problem-solving capabilities, but large-scale deployment in actual operations remains sparse due to its inherent limitations (such as hallucinations) and the vast array of organizational and institutional barriers to its effective use in different domains.

4. AUTOMATING TASKS AND JOBS

Debates regarding the business implications of AI have advanced on several distinct levels. On one level, a significant body of research focuses on AI’s ability to automate a broad array of tasks and jobs. On another level, we can draw valuable insights from how past technologies have transformed

organizations and institutions. These lessons are essential for unleashing the full potential of AI and provide crucial guidance for its adoption and exploitation. On a third level, we need to engage in future-thinking and “future scaping”, using our imagination to conceive of the changes that AI can bring about in our economies and society. Arguably, it is the failure of imagination that is holding us back more than anything else.

Many extant studies have examined how AI can automate tasks and jobs, but so far, most projections based on such studies have not materialized. On November 24, 2016, the Godfather of AI, Geoffrey Hinton, famously argued that “People should stop training radiologists now. It’s just completely obvious that within five years, deep learning is going to do better than radiologists.”¹⁰ However, seven years after that remark, there is still a 29% shortfall of radiologists and 15% shortfall of clinical oncologists in the NHS in the U.K.; and the six-week waiting list for CT and MRI scans is still increasing, not decreasing.¹¹ The experience in the U.S. is similar.

The reasons for such wildly off the mark predictions are complex, but a cursory look at the 30 or so tasks that radiologists routinely perform shows that no more than a handful can be automated by AI, requiring human radiologists to continue to perform the majority of the tasks (Table A1, in the Appendix).¹² In addition, radiologists must also consider other important issues such as technology skills, specialist tools, and they must work closely with other specialists and professionals and navigate the complex organizational and regulatory environments for health services. In 2024, Nvidia CEO Jensen Huang made similar predictions that AI would soon eliminate the need for coders. This prediction is probably not going to age well.¹³

It is important to note that while research reports and media headlines often emphasize that AI systems outperform humans on various benchmarks, from reading comprehension to professional exams, this does not necessarily mean AI surpasses human capabilities in the tasks these benchmarks represent. AI performance on benchmarks often fails to accurately predict how it will perform in real-world scenarios.

Additionally, even for the tasks and roles that AI can fully automate, the transition period may be lengthy. Carl Benedikt Frey of the University of Oxford uses the “lampighter” as

¹⁰ <https://tinyurl.com/3ctvhfea>

¹¹ <https://tinyurl.com/vzrrt8zh>

¹² <https://tinyurl.com/32z673yb>

¹³ It is important to note that coding is becoming increasingly democratized through AI. In time, anyone will be able to generate working code or software applications with AI assistance, even without coding knowledge. This shift will push specialist coders into more niche areas, though the full evolution of this process may take considerable time.

an illustrative example. Initially, when streetlights were gas-powered, individuals were employed to light each lamp at dusk using a flaming wick on a long pole.¹⁴ With the advent of electric bulbs, lamplighters continued to work, manually switching on each light. However, as cities implemented block-wide switches and later, timers and light sensors, manual intervention became obsolete. Frey suggests that AI might undergo a similar evolution.

It is also worth noting that many studies, including those from Frey and his colleagues, also suggest that additional jobs will be created – either because greater efficiency leads to increased demand (Jevons Paradox) or entirely new jobs are required. This aspect is often missing from many of the debates about AI's impact on employment. We currently see minimal job displacement, and this could give a false sense of security. Nevertheless, the full impact could take many years, if not decades, to unfold, giving people time to adapt.

5. TRANSFORMING ORGANIZATIONS AND INSTITUTIONS

To understand the full economic and social impact of AI, we need to look beyond the automation of tasks and jobs to explore how AI transforms organizations and institutions. History suggests that if AI changes our lives, it will not occur overnight but more likely to unfold over decades rather than years. At first, these changes are likely to be gradual, integrating into existing organizational settings and lifestyles first, gradually transforming them through experimentations, disruptions, and generational transition.

However, we must also question whether future projections based on past lessons will prove accurate in this case. As Ernest Hemingway famously remarked when asked how he went bankrupt: “Gradually, then suddenly”. The risk with AI is that its impact may appear slow to develop, only to arrive abruptly and dramatically. Digitalization has already demonstrated how quickly and unexpectedly such sweeping changes can happen, highlighting the need to address these challenges now.

Considering incremental versus transformative change, Clayton Christensen's “The Innovator's Dilemma” provides an important lesson for understanding how disruptive technologies, like GenAI, can reshape industries.¹⁵ In this context, LLMs and increasingly multimodal AI represent

potentially disruptive innovations that may initially seem limited but are rapidly evolving in terms of price, availability, and capability. Traditional firms face a dilemma: they must balance serving existing customers and markets with the need to adapt to these emerging technologies. The rapid advancement of GenAI is likely to outpace market expectations – no matter how skeptical people are – potentially rendering some existing products and services obsolete.

As a result, established companies may struggle to adapt their existing value networks and capabilities to the new AI paradigm. This creates opportunities for more agile, AI-native firms to gain footholds in emerging markets. To navigate this landscape successfully, traditional firms will need to reskill their workforce, reconsider their business models, and adopt more flexible, experimental approaches to innovation. The democratization of AI capabilities may erode some competitive advantages of larger organizations, forcing them to find new ways to create and capture value in an AI-driven world. One consequence is that, instead of a single “big bang”, numerous small, incremental (and radical) changes will slowly and cumulatively reshape how we live and cooperate over time – a process that may continue for decades and span generations.

The history of technological advancements during the Industrial and Digital Revolutions does, though, offer several valuable insights for understanding the transformative potential of AI. The Industrial Revolution began in the late 18th century but took over a century to realize its impact. Early adopters gained some competitive advantages, yet the widespread transformation of industries occurred incrementally. Similarly, the introduction of computers, the internet, and mobile reshaped businesses and consumer behavior, but again, this was a gradual transition that spanned decades.

5.1 Industrial Revolution

During the Industrial Revolution (1760-1840), textiles were the first industry to see factories filled with machines that automated many tasks. This shift was powered by new energy sources like coal and steam, leading to the rise of large industrial cities and rapid urbanization. The advent of mechanization, such as the spinning jenny and power loom, alongside the development of factories, centralized production and drastically increased efficiency. However, the transition unfolded over many decades, with gradual adoption and incremental advancements, eventually revolutionizing how goods were produced and reshaping society.

¹⁴ <https://tinyurl.com/y2ta2j2z>

¹⁵ <https://tinyurl.com/4s6hnu9r>

The shift from steam to electric power was also gradual. In 1879, Thomas Edison famously unveiled the electric light bulb, yet by 1900, only 3% of U.S. households had electricity, reaching 50% only by 1920 after over 40 years. The adoption of electric power in factories was even slower. This highlights the slow and gradual integration of new technologies into daily life and industry. Importantly, the real benefits of electrification did not come from reduced costs from cheaper power, and there were significant transitional costs involved. Unlike steam power, electrification facilitated the use of distributed, fractionalized power by allowing electric motors to be mounted on individual machines. This enabled a shift from the traditionally vertical, multi-story, cramped factory designs centrally powered by steam to more efficient horizontal layouts. Initially, electrification was adopted in emerging industries of the time, such as tobacco and transport equipment, rather than in incumbent industries such as textiles. This is similar to the adoption of AI so far, which has shown parallel trends, initially being embraced in emerging areas such as search, e-commerce, social networking, and online streaming, rather than more traditional industries in manufacturing and services. However, it is important to note that the adoption of AI has also been focused on automating and augmenting existing work-related tasks and processes, particularly in back-office functions, such as IT, finance, legal, marketing and HR, exactly as when electricity was introduced to factories. It is only when organizations realize that they need to change the fundamentals of work, tasks, and entire functions and business models that we will start to see the real benefits of AI.

5.2 Digital Revolution

The patterns observed during the Digital Revolution were similar. Mainframe computers were first introduced in the late 1950s and 1960s, progressing to distributed computing in the 1980s and 1990s with the advent of mini-computers, PCs, and distributed architecture. The consumer Internet was commercialized in early 1990s, followed by the mobile internet in the late 1990s and 2000s. This expansion significantly accelerated with the advent of smartphones and 3G and 4G mobile networks, starting with devices like Blackberries and Palms, and iPhones from 2007. However, the productivity paradox, articulated by Robert Solow as “you can see the computer age everywhere but in the productivity statistics,”¹⁶ persisted throughout this period, except for 1994 to 2005. Since the late 2000s, productivity has stagnated again, a trend that continues to this day.

5.3 The AI Revolution

AI's impact on organizations is likely to follow a similar trajectory. Currently, AI has made significant progress in powering complex systems like Google search, Amazon and Alibaba recommendations, and Uber and Didi matching. These advancements demonstrate its capabilities in automating tasks, streamlining processes, enhancing user experiences, and enabling new business models. However, as AI expanded into more traditional sectors, it encountered significant hurdles. For instance, despite decades of anticipation for self-driving cars since the 1980s, successful implementations have been elusive. Tesla has repeatedly postponed the debut of fully autonomous vehicles, while companies such as Apple abandoned their autonomous driving projects, and Uber dissolved its self-driving unit. In retail, ventures like Amazon Go, introduced as cashier-less stores in 2018, have not achieved broad acceptance, casting doubt on their feasibility. Similarly, initiatives like Freshippo (Hema), central to Alibaba's “New retail” strategy, have only made modest contributions, highlighting the complexities and uncertainties surrounding AI's integration into these sectors. Such developments call for significant organizational and institutional changes, which require time to evolve and get right.

These examples primarily come from B2C organizations, shaped by consumer attitudes and behaviors towards AI. In contrast, B2B and other areas of B2C business may present a different narrative. The challenge is not just about having the technological capability to drive substantial change, but also about successfully implementing AI at scale and transforming existing cultures and processes.

From the steam engine and electricity to the computer and mobile phone, integrating new technologies into business operations has historically taken decades. While this demonstrates the need for business leaders to develop a long-term strategy for AI integration, they must also navigate the delicate balance between managing expectations of a gradual transformation and the possibility of rapid change. The challenge lies in determining where this balance is and finding the right language to communicate it effectively.

One approach is to adopt Daniel Kahneman's concept of “Thinking fast and slow” – some tasks in business need to be done quickly, while others can progress more slowly. Focusing solely on speed or caution is not the right solution. Doing some

¹⁶ <https://tinyurl.com/6nr6w7dm>

¹⁷ <https://tinyurl.com/yz8yk5dp>

things quickly enables the slower elements to be improved, as the business learns and iterates. Recognizing and balancing these nuances is essential for achieving sustainable success.

6. MANAGING THE TRANSITION

Technological advancements often happen rapidly, but organizational and institutional changes tend to occur more slowly and are iterative and fraught with complexities, involving adjusting regulatory frameworks and societal norms and overcoming resistance to change. As Clay Shirky noted, “Institutions will try to preserve the problem to which they are the solution.”¹⁸ Successfully managing the transition to new technologies and new organizational designs requires navigating these complexities. Simply automating tasks and jobs will not be sufficient to unlock AI’s full potential. Instead, AI should be used to reimagine operational processes and business models, and the wider institutional environment. Regulatory frameworks, educational systems, and ethical standards will need to be updated to accommodate the rapid development of AI. This includes addressing issues such as data privacy, the ethical use of AI, intellectual property rights, transparency in AI-driven decisions, individual protection from algorithmic bias, and importantly, ensuring that AI advancements benefit society. Such organizational and institutional changes tend to be much slower than the pace of technological developments.

The transition to new technologies and institutions is rarely cost-free. Our research with senior business leaders from the U.S., Europe, and China shows that strategic initiatives often fail, not because the ideas are intrinsically flawed, but due to leadership failure to effectively manage the transition to new technologies, organizational designs, and business models.¹⁹ There is typically too much focus on the technology and its promise and not enough on all the other elements necessary for successful transformation – especially people and change.

The case of digital health illustrates the gap between their promised potential to reduce costs and increase efficiency in healthcare, and the reality of their slow deployment in real operations. Despite significant investment, integrating new digital technologies into healthcare systems has been fraught with challenges around the world. Usability issues, inadequate training, concerns over patient data security, and necessary changes to medical or administrative procedures are common hurdles. Overcoming these obstacles requires comprehensive

“

To understand the full economic and social impact of AI, we need to look beyond the automation of tasks and jobs to explore how AI transforms organizations and institutions. ”

planning, active engagement with key stakeholders, significant new resources, and iterative testing to align new technology with the needs of stakeholders.

Moreover, these changes must be implemented while the existing systems are fully operational, often under conditions where staff are already facing high pressure and have little capacity to adopt new processes. The frustrations are palpable, as illustrated by the CEO of a major NHS hospital who exclaimed in an interview with one of the authors: “I am going to punch the next son of a b**** who tells me his technology is going to save me money!” Managing the transition means maintaining the existing operation while finding additional resources to support the new processes, and it will lead to increased overall cost in the short to medium term – and the long-term savings are by no means guaranteed – and the process often outlasts the tenure of many senior leaders in these organizations. This vividly highlights the practical challenges of integrating new technologies into established healthcare infrastructures. Deploying AI is unlikely to escape such constraints, and the transition process must be effectively managed.

7. IMAGINING THE FUTURE – STRATEGIC VISION FOR AI TRANSFORMATION

While using AI to automate tasks and jobs may yield short-term gains, its true transformative potential lies in its ability to fundamentally redesign how organizations operate and shape their environments. However, AI itself cannot redesign organizational structures; it is up to people to rethink operating models and business processes based on AI’s capabilities. Managing this transition will be challenging, as many large

¹⁸ <https://tinyurl.com/d9ztnkrd>

¹⁹ Li, F., 2020, “Leading digital transformation: three emerging approaches for managing the transition,” *International Journal of Operations and Production Management* 40:6, 809-817

organizations may resist change, particularly if their industries remain profitable. As a result, radical innovations are more likely to emerge in startups within developing sectors, setting the stage for future industry disruptions and reshaping traditional landscapes. These shifts will not occur overnight; as progress will likely be uneven, with both small and large leaps forward, making the transition unpredictable and difficult to navigate.

For senior business leaders, the traditional linear approach to strategy development and implementation is no longer fit for purpose. Instead, strategy formulation and execution must become iterative and intertwined, especially when both the path and destination for the organization may undergo frequent adjustments. This approach allows strategies to evolve in real-time, informed by ongoing execution and feedback. By adopting such iterative processes, organizations can continuously adapt to new intelligence and align with shifting goals and market conditions.²⁰

Learning from historical technological transformations is crucial for successfully integrating AI into their strategic planning and operational processes. Business leaders must look beyond immediate efficiencies gained from AI automation and prepare for broader organizational and institutional changes. There will be many challenges that we cannot currently foresee, but by developing a deep, nuanced understanding of AI's potential, rooted in both historical insights and emerging realities, leaders will be better equipped to navigate and succeed in the rapidly evolving AI era.

8. CONCLUSION: BEYOND TRANSFORMING BUSINESS AND INSTITUTIONS

The advent of AI has also ignited discussions about creativity and innovation, particularly concerning AI's role in activities traditionally seen as distinctly human, such as art. Contrary to common perceptions about AI's capabilities, art is fundamentally about intent and communication, serving as a medium to evoke emotions and convey messages. Controversially, a renowned artist remarked when asked about how AI will likely transform art: "Art is exactly the opposite of AI. Art emerges from intent, from a desire to express something, to communicate something, to make someone else feel something. Art, in all its forms, is primarily about communication, not just a collection of colors or words. If you view AI-generated art as competition, it might be time to reconsider the reasons behind your own writing or painting."

AI will also have a significant impact on society. It is crucial to remember that the introduction of AI into organizations should benefit the entire society, not just a privileged few. The policy implications have not been fully understood. As one senior policymaker noted, "It's about power dynamics and how we choose to organize ourselves as a society. Until we find a better way to manage our resources, every change will adversely affect the unprivileged."

These issues are fundamentally important in the AI-driven era, and further research is needed to understand the long-term business and societal implications. While AI offers significant opportunities for efficiency, innovation, and transformation, it also challenges traditional notions of creativity and raises complex societal concerns. Balancing technological progress with business goals, ethical considerations, and equitable access will be crucial. Policymakers, organizational leaders, and individuals must work together to address these challenges and ensure AI enhances human potential without exacerbating inequality. The strategies business leaders choose will play a key role in shaping the future of work and society.

²⁰ Li, F., 2022, "Sustainable competitive advantages via temporary advantages: insights from the competition between American and Chinese digital platforms in China," *British Journal of Management*, 33:4, 2009-2032

APPENDIX

Table A1: Job duties for radiologists

1.	Prepare comprehensive interpretive reports of findings.
2.	Perform or interpret the outcomes of diagnostic imaging procedures including magnetic resonance imaging (MRI), computer tomography (CT), positron emission tomography (PET), nuclear cardiology treadmill studies, mammography, or ultrasound.
3.	Document the performance, interpretation, or outcomes of all procedures performed.
4.	Communicate examination results or diagnostic information to referring physicians, patients, or families.
5.	Obtain patients' histories from electronic records, patient interviews, dictated reports, or by communicating with referring clinicians.
6.	Review or transmit images and information using picture archiving or communications systems.
7.	Confer with medical professionals regarding image-based diagnoses.
8.	Recognize or treat complications during and after procedures, including blood pressure problems, pain, oversedation, or bleeding.
9.	Develop or monitor procedures to ensure adequate quality control of images.
10.	Provide counseling to radiologic patients to explain the processes, risks, benefits, or alternative treatments.
11.	Establish or enforce standards for protection of patients or personnel.
12.	Coordinate radiological services with other medical activities.
13.	Instruct radiologic staff in desired techniques, positions, or projections.
14.	Participate in continuing education activities to maintain and develop expertise.
15.	Participate in quality improvement activities including discussions of areas where risk of error is high.
16.	Perform interventional procedures such as image-guided biopsy, percutaneous transluminal angioplasty, transhepatic biliary drainage, or nephrostomy catheter placement.
17.	Develop treatment plans for radiology patients.
18.	Administer radioisotopes to clinical patients or research subjects.
19.	Advise other physicians of the clinical indications, limitations, assessments, or risks of diagnostic and therapeutic applications of radioactive materials.
20.	Calculate, measure, or prepare radioisotope dosages.
21.	Check and approve the quality of diagnostic images before patients are discharged.
22.	Compare nuclear medicine procedures with other types of procedures, such as computed tomography, ultrasonography, nuclear magnetic resonance imaging, and angiography.
23.	Direct nuclear medicine technologists or technicians regarding desired dosages, techniques, positions, and projections.
24.	Establish and enforce radiation protection standards for patients and staff.
25.	Formulate plans and procedures for nuclear medicine departments.
26.	Monitor handling of radioactive materials to ensure that established procedures are followed.
27.	Prescribe radionuclides and dosages to be administered to individual patients.
28.	Review procedure requests and patients' medical histories to determine applicability of procedures and radioisotopes to be used.
29.	Teach nuclear medicine, diagnostic radiology, or other specialties at graduate educational level.
30.	Test dosage evaluation instruments and survey meters to ensure they are operating properly.

Source: O.Net Online (<https://tinyurl.com/hmc4n3kd>)

THE CHALLENGES OF AI AND GenAI USE IN THE PUBLIC SECTOR

ALBERT SANCHEZ-GRAELLS | Professor of Economic Law, University of Bristol Law School

ABSTRACT

In this short paper, I reflect on the challenges that the public sector faces in adopting artificial intelligence (AI), and generative AI (GenAI) in particular. Despite the increasing pressure on public sector organizations to deploy AI and GenAI to cut costs, this stage of public sector digitalization remains fraught with difficulties. I stress in particular the challenges that arise from the two-tier complexities of: first, designing appropriate use cases and ensuring AI and GenAI are not used for other purposes and, second, successfully acquiring AI and GenAI for the public sector.

1. INTRODUCTION

Given the progressive (and at times sudden) mainstreaming of artificial intelligence (AI), and generative AI (GenAI) in particular, across all industries, it seems unavoidable for public sector organizations to seek to harness the opportunities they bring.

Crucially, AI and GenAI are being targeted as key sources of savings for the public sector. For example, a recent report estimated that, in the U.K., “greater use of AI to support the completion of routine tasks and administration in the public sector could create over £12 billion in savings for the public sector by 2030. By 2035 greater use of AI could save the UK’s public sector £17 billion.”¹ Similar estimates and projections abound for almost all jurisdictions. In a context of fiscal challenges and macroeconomic uncertainty, the promise of savings of this scale cannot be ignored by governments. And, in fact, some governments are putting significant hopes on these technologies to plug funding gaps and/or modernize their public services,² as well as exploring ways in which the public sector can act as an incubator or living lab for tech start-ups.

One would be forgiven for harboring doubts about the feasibility of sticking to this sort of timeline and the viability of organizational and cultural changes of the magnitude required to achieve such savings. They are similar to those required to start to tackle climate change and decarbonization – and the track record in that area is not very encouraging at all. Whether the (seemingly) high-powered public finance incentives involved in the AI context will make a difference is anyone’s guess.

Setting that aside for now, and glossing over the fact that numerous claims on the potential of AI and GenAI (as well as other digital technologies) are but new forms of snake oil,³ I am interested in reflecting on the challenges faced by public sector organizations seeking to deploy AI and GenAI. From my point of view, and after conducting extensive research in the area of public sector digitalization,⁴ there are noticeable challenges that arise from the two-tier complexities of: first, designing appropriate use cases and ensuring AI and GenAI are not used for other purposes and, second, successfully acquiring AI and GenAI for the public sector.

¹ Microsoft/Public First, 2024, “Unlocking the UK’s AI potential: harnessing AI for economic growth,” May, 32, <https://tinyurl.com/4t2ay3j4>

² In the case of the U.K., see Department for Science, Innovation & Technology, Artificial Intelligence (AI) Opportunities Action Plan: terms of reference (July 26, 2024), <https://tinyurl.com/2yb7t48n>

³ Narayanan, A., and S. Kapoor, 2024, AI snake oil: what artificial intelligence can do, what it can’t, and how to tell the difference, Princeton University Press

⁴ Sanchez-Graells, A., 2024, Digital technologies and public procurement. Gatekeeping and experimentation in digital public governance, Oxford University Press

2. HOW TO IDENTIFY “GOOD” USE CASES AND AVOID “BAD” DEPLOYMENTS

Identifying appropriate use cases for AI and GenAI is a challenge for the public sector.

At one level, there are significant issues with the data and IT/software architectures of the public sector that make it hard to “plug AI” on top of them. Limited access to structured historical data can make it difficult to train or fine-tune AI and GenAI models for deployment in public sector specific contexts. Worse still, historical data that embeds biases and discrimination may be impossible to “clean”, and any application of AI based on such data would perpetuate and amplify those historical sources of injustice. It can also be difficult to find ways to integrate AI and GenAI provided over cloud infrastructures with some of the legacy systems still running in the public sector.

However, as far as I can see, these are “technical” challenges and not too different from those faced in other sectors, such as the financial services industry. Given adequate resourcing (and this is a big if, both in terms of total funding but also, crucially, in terms of the public sector digital workforce) they can probably be overcome.

At another, deeper level, the public sector faces significant challenges identifying “good” use cases from the perspective of the duties it owes citizens, and broader concerns with core and fundamental values, as well as legal rights. Just because an administrative process “can”, for example, be automated through AI solutions or “elevated” with GenAI, this clearly does not mean it “should” be. There is a rapidly stacking pile of evidence, across jurisdictions such as Australia⁵ or the Netherlands,⁶ that shows that use cases that may make sense from the narrow perspective of procedural optimization within the public administration (even through forms of automation or algorithmic decision making not involving AI) carry excessive risks and are unlikely to be acceptable to citizens once their operation and effects are uncovered.

This concerns the use of AI or GenAI for citizen-facing services such as the administration of benefits, tax, or the social services, criminal and prison systems. An interesting tension

here is that it seems to be the case that some of the potential big gains of deploying AI and GenAI are linked to mass or population-wide services. However, these are also the services where the deployment of AI or GenAI will be most likely to carry excessive risks.⁷ This poses a particular challenge for the public sector because the effect of failed or perverse AI deployments on citizens’ trust is very different from, say, the reputational effects of similar failures in the private sector. Moreover, the legal risks associated with such AI use cases are also rather particular.

To be sure, the emerging stories of failure in the deployment of AI, and GenAI in particular, in the private and voluntary sectors serve as a cautionary tale for the public sector. Recent months have seen rushed deployments of GenAI result in damages awards against Air Canada where the “hallucination” of its chatbot inaccurately explained the airline’s bereavement policy,⁸ or the reputationally damaging short-lived deployment of a chatbot launched by the National Eating Disorders Association in the U.S. to teach people experiencing eating disorders coping skills, when it became evident that the AI was offering users advice for weight loss instead.⁹ These and other cases show that much more care has to be exercised in the deployment of AI and GenAI where the stakes are high. And, by definition, the stakes will tend to be much higher in (involuntary) interactions with the public sector than in (commercial or nonprofit) interactions with the private and nonprofit sectors.

This restricts most of the relatively less controversial uses of AI to highly technical fields, such as healthcare (in jurisdictions where this is a public service), where AI can more readily be used as a tool to support or enhance processes in narrowly defined application domains (such as radiography). In these cases, deploying AI and GenAI will still face the “procurement challenge” discussed below. In all other circumstances, the public sector needs to approach the identification of “good” use cases with caution and find effective strategies to engage relevant stakeholders, mitigate all relevant risks, and ensure sufficient “social buy in”. Although there are emerging frameworks to support these assessments and decision making processes,¹⁰ they are still in their early stages and will require significant effort in their implementation.

⁵ Royal Commission into the Robodebt Scheme, 2023, “Final report,” July 7, <https://tinyurl.com/mrx6c42j>

⁶ Heikkilä, M., 2022, “Dutch scandal serves as a warning for Europe over risks of using algorithms,” Politico, March 29, <https://tinyurl.com/4ckjbxky>

⁷ Sanchez-Graells, A., 2024, “Resh(AI)ping good administration: addressing the mass effects of public sector digitalization,” *Laws* 13:1, 9, <https://tinyurl.com/mrxkr3xd>

⁸ Belanger, A., 2024, “Air Canada has to honor a refund policy its chatbot made up,” *Wired*, February 17, <https://tinyurl.com/273scqpb>

⁹ Van Amburg, J., 2023, “AI is now a destructive steward of diet culture,” *Well + Good*, August 17, <https://tinyurl.com/485kfkej>

¹⁰ See, e.g., IEEE, 2021, “Standard for the procurement of artificial intelligence and automated decision systems (in progress),” <https://tinyurl.com/3ywehywh>. See also, Waters, G., and C, Miller, 2024, “5 ways to strengthen the AI acquisition process,” *IEEE Spectrum*, March 26, <https://tinyurl.com/yk6478yj>

In general, this does not seem to necessarily dissuade public sector leaders from seeking to use AI and GenAI, and there are clear indications that some sectors, such as education, are being targeted for AI-related investments¹¹ despite the absence of evidence (or a clear ethical and legal framework) on the effects of AI and GenAI exposure on schoolchildren and students¹² – but it tends to push those pilots and deployments behind a curtain of opacity and secrecy. In most jurisdictions, there have been very limited advances in ensuring adequate transparency and accountability for public sector AI use. Although there is an emerging trend to strengthen governance of the use of AI in the public sector – such as with the U.S. Executive Order on AI,¹³ some aspects of the E.U. AI Act,¹⁴ or the very recent Framework Convention on artificial intelligence and human rights, democracy, and the rule of law¹⁵ – there is still a long way to go to ensure adequate and effective implementation. It will be a few years until the regulatory and governance frameworks required by these emerging international and domestic norms are fully embedded.

This leads to a final related challenge concerning the “unauthorized” or “unregulated” use of AI and GenAI in the public sector. In many cases, public sector organizations will not yet have adopted AI or GenAI solutions that “could” be deployed in their activities. This places those organizations in a difficult position if individuals within them make use of those technologies, or if incumbent IT vendors embed AI in ways that are not visible or traceable for the organization, or from which it cannot (technically) opt out. Even if organizations formally ban the use of those technologies (e.g., by preventing access through organization-administered IT), or issue guidelines on what they consider appropriate use,¹⁶ they need to come up with additional measures to avoid individuals working around such bans or technical or organizational constraints (e.g., by using GenAI on their personal devices and then forwarding the relevant outputs to their work email for subsequent use within the “permitted” official workflow). They also need to develop ways to audit (inadvertent) AI embeddings in increasingly complex digital supply chains. To some extent, AI and GenAI use “in” the public sector is distinct from its use

“by” the public sector and this requires organizations to align individual and vendor behavior with their official position and legal obligations.

3. HOW TO SUCCESSFULLY PROCURE AI AND GenAI

As mentioned above, where a public sector organization finds a “good” and viable use case, there is still the challenge of acquiring (or procuring) the technology – as very few organizations will be in a position to develop it in-house. AI procurement, and GenAI in particular, poses a particular challenge, even compared to that of other types of complex (software) systems because, except for “off-the-shelf” AI solutions, it poses technical and contextual risks that we are yet to fully understand, and because public buyers cannot (yet) rely on traditional de-risking tools – which leaves them exposed to regulatory and commercial capture. This challenge breaks down into many different dimensions.

Public buyers will have a difficulty defining the type of AI (or GenAI) solution they seek to acquire. This will be difficult because they may not want to (or be able to) prescribe a specific solution in a quickly-changing marketplace, or because there may be different technical ways of achieving a similar functionality and the procurement process will need to tease out the overall preferable approach once trade-offs between technical, financial, and governance implications are clear. It can also be difficult because the public buyer may have gaps in its digital capabilities or market research and may need to use the tendering process to get a better view of what the market can offer (that is, to gauge the “state-of-the-art”).

Public buyers will also face issues setting technical specifications and organizational arrangements in a context where there is no clear consensus on what these need to entail and where work by international standardization bodies is still in progress. Moreover, some of the parameters that public buyers will need to specify, such as the accuracy, robustness (including cybersecurity), and explainability of the

¹¹ See, for example, in the U.K., Department for Education and Department for Science, Innovation & Technology, Research on public attitudes towards the use of AI in education (28 August 2024), <https://tinyurl.com/ykafk2hn>

¹² See, for example, Ali, O., P. A. Murray, M. Momin, Y. K. Dwivedi, and T. Malik, 2024, “The effects of artificial intelligence applications in educational settings: Challenges and strategies,” *Technological Forecasting and Social Change*, 199: 123076, <https://tinyurl.com/bdrx353y>

¹³ Executive Order 14110 on safe, secure, and trustworthy development and use of artificial intelligence of October 30, 2023, <https://tinyurl.com/3c7apx5d>

¹⁴ Regulation (EU) 2024/1689 of the European Parliament and of the Council of June 13, 2024 laying down harmonized rules on artificial intelligence, <https://tinyurl.com/4e3s3h23>

¹⁵ Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, CETS No. 225, <https://tinyurl.com/99pvrz7m>. The U.S., E.U., and U.K. all signed the treaty on the first day it was open for signature.

¹⁶ See, for example, for the U.K., Cabinet Office and Central Data Office, Guidance to civil servants on use of generative AI (January 29, 2024), <https://tinyurl.com/3crbwp6f>

AI and GenAI systems are very much in flux and under ongoing research. In this context, it can be difficult to run a procurement process with the required level of predictability and to ensure a level playing field in the conduct of negotiations and technical dialogs.

Public buyers will also have difficulties coming up with award criteria and structured ways to assess offers that could vary across a wide range on cost and quality (e.g., capability or environmental impact), as well as ensure that the terms and conditions that get embedded in the contract do not generate unforeseeable costs or carry undesirable implications (such as lock-in). Given the different strategies used by AI and GenAI companies to monetize their products, this can be a particular challenge where there is no industry standard.

This is linked to the difficulty in assessing claims of compliance with whichever technical specifications are used, or to assess the adequacy of “state-of-the-art” offers where the public buyer does not have the technical competency or capability to, for example, directly test the AI or GenAI. Alternative approaches, such as third-party certification or assurance are also not yet well-developed and, in the same way that there are no generally accepted industry technical standards, there are no generally accepted audit techniques and standards either. This places public buyers in a difficult position because requiring third-party certification or audit can well displace the focus of the market for lemons (from the AI solution to the auditor and its methods), but not solve the problem.

Relatedly, public buyers will find it difficult to impose their terms and conditions and to negotiate specific issues where there is an imbalance of power with the tech vendors (or a Big Tech company embedded along the supply chain, such as when “start-up offers” are built on “off-the-shelf” platforms or components controlled by bigger players). Public buyers cannot (just) hope to have market power to an extent that allows them to dictate the terms of the relevant contract.

There are further complications, but these should suffice to show that procuring AI will be challenging and that public buyers will not have ready access to de-risking tools they can usually use in other contexts, such as requirements to comply with technical standards, audit and certification, or “take it or leave it” tendering and contract conditions.

4. CONCLUSION: A DIFFERENT APPROACH

Given the significant challenges in identifying adequate good cases for AI and GenAI in the public sector and to successfully procure the technology, I would argue that a different approach is required. The emerging strategy of self-regulation by the public sector in choice of use cases and the attempt to use contract-based regulation to govern the acquisition and deployment of AI and GenAI are unlikely to result in robust processes for public sector digitalization capable of protecting the public interest and fostering citizen trust.

In my view, governments that want to take the opportunities of AI and GenAI seriously will have to start by putting an adequate legislative and regulatory framework in place. My specific proposal¹⁷ is for a dedicated regulator in charge of a system of licensing of public sector AI use not too dissimilar in its foundations to the food and drug regulators in Western jurisdictions. To put it differently, jurisdictions need to quickly move away from the light-touch regulatory approach that is becoming the global standard. This will require investment in this needed additional layer of administration, as well as in upskilling the public sector on digital issues. However, this investment is required to ensure that the public sector is in the driving seat in the process of digitalization and that it brings citizens with it in a safe and trustworthy way.

The alternative perhaps looks bleak. A jurisdiction that pushed ahead with the deployment of AI and GenAI in the public sector solely in pursuit of (medium term) financial savings would likely be betting on a losing strategy and one that could well leave it locked into technologies and tech vendors over which it has limited effective regulatory levers, and with waning support and trust from its citizens after repeated scandals and instances of discrimination and human rights breaches. I think it is no exaggeration to say that the window of opportunity to put the fundamentals in place to steer the digitalization of the public sector is relatively narrow. And this is also something the digital transition has in common with the much urgently required green transition. I for one hope to see swift regulatory and legislative change and for the dominating trend of decision making in the AI and GenAI context to be brought back to the public sphere and away from Big Tech vendors.

¹⁷ Sanchez-Graells, A., 2024, ‘Responsibly buying artificial intelligence: a “regulatory hallucination”’ Current Legal Problems, cuae003, <https://tinyurl.com/4xpzs28n>

AI SAFETY AND THE VALUE PRESERVATION IMPERATIVE

SEAN LYONS | Author of Corporate Defense and the Value Preservation Imperative: Bulletproof Your Corporate Defense Program

ABSTRACT

Global artificial intelligence (AI) safety is critical to defending against the potential downside of AI technology (from routine to existential risks) and needs to be prioritized accordingly. Our global leaders have a duty of care to safeguard against the potential damage of this impending AI value destruction and that will require a much higher, more robust, and more mature level of AI safety due diligence than is currently on display. Dynamic developments in AI mean that the normal order of things no longer applies, and that going forward effective AI safety will require superior levels of guardianship, stewardship, and leadership.

1. INTRODUCTION: THE NEED FOR GLOBAL AI SAFETY STANDARDS AND PRACTICES

AI technology, as it continues to evolve (i.e., narrow AI, general AI, interactive AI, etc.), is likely to contribute to the creation, preservation, and destruction of stakeholder value. The recent increase in the proliferation of AI clearly presents extraordinary benefits and opportunities for both the corporate world and for humanity. Exceptional rewards are, however, also accompanied by equally exceptional risks. The dynamic nature of these new AI technologies means that the digital age has become increasingly complicated and is leading to a level of complexity that humankind is already struggling to fully comprehend.

The challenge presented by AI is a global challenge and one which requires a global approach and global solutions. Due to the pervasive nature of AI technology, it has the potential to have both positive and negative impacts at organizational, national, international, and global levels. Humanity, therefore, needs to ensure that appropriate safeguards and guardrails are in place and operating effectively at all levels. Addressing this matter is by no means an easy task, but it is one that needs to be viewed as a mandatory obligation. As the concept of AI safety is still in its relative infancy, there is

currently no single, unified, globally agreed upon approach to collectively safeguard stakeholder AI value. Currently, AI safety developments appear to be organic rather than systematic in nature, with different countries and regions adopting varying frameworks, regulations, and priorities. Consequently, in recent years serious safety concerns have been publicly expressed by AI experts, researchers, and backers [FLI (2023)].

This paper is focused on applying the corporate defense management (CDM) philosophy and principles [Lyons (2016)] to the AI safety challenge to provide organizations with a high-level roadmap to help address these AI safety concerns, and to help ensure that appropriate safeguards and guardrails are in place.

1.1 The upside of AI – potential rewards

In terms of the potential upside, digital and smart technologies are already pervasive and AI in its many forms (i.e., machine learning, natural language processing, computer vision, etc.) has the potential to leverage from this to add significant value, to make enormous contributions, and to create long-term positive impacts for society, the economy, and the environment. It has the potential to solve complex problems and create opportunities that benefit and reward all human beings and their ecosystems [OSTP (2022)].

1.2 The downside of AI – potential risks

Unfortunately, AI systems also have the potential for extreme downside, and to cause an unimaginable level of harm and damage to human ecosystems (i.e., business, society, and planet). Its potential for destruction stems from the dangers associated with the risks, threats, and hazards associated with AI [NIST (2024)] and these could manifest themselves in the form of not only their initial impact but also their potential collateral damage.

1.3 AI dangers and collateral damage

Examples of the dangers posed by AI technology relate to the potential negative impact of the following scenarios [Lyons (2024a)]:

- Environmental sustainability and destruction:** AI technology is capable of consuming massive amounts of both energy and water, which has the potential to detrimentally impact on the environment. A lack of transparent disclosure on environmental footprints, practices, and impacts can have a negative and destructive impact on environmental sustainability. Unregulated AI can potentially contribute to global warming through its greenhouse gas emissions, result in energy shortages in residential power supply due to the impact of its energy intensive nature on our national grids, and negatively impact on water security (and pollution) due to the industry's need for water to cool its physical machines [Mazzucato (2024)].
- Misuse and abuse:** AI technologies can be misused and abused for all sorts of malicious purposes with potentially catastrophic results. They can be used for deception, to shape perceptions, or to spread propaganda. AI generated deepfake videos can be used to spread false or misleading information, or to damage reputations. Other sophisticated techniques could be used to spread misinformation and be used in targeted disinformation campaigns to manipulate public opinion, undermine democratic processes (e.g., elections and referendums), and destabilize social cohesion (e.g., polarization and radicalization).
- Privacy, criminality, and discrimination:** AI powered surveillance, such as facial recognition, can be intentionally used to invade people's privacy. AI technologies can help in the exploitation of vulnerabilities in computer systems and can be applied for criminal purposes, such as committing fraud or the theft of

“

The paradox of AI is that eventually only AI technology will have the capability to manage the complexity of AI technology.

”

sensitive data (including intellectual property). They can be used for harmful purposes, such as cyberattacks (including cyberterrorism), and to disrupt or damage critical infrastructure. In areas such as healthcare, employment, and the criminal justice system, AI bias can lead to discrimination against certain groups of people based on their race, gender, or other protected characteristics. It could even create new forms of discrimination potentially undermining democratic freedoms and human rights.

- Job displacement and societal impact:** As AI related technologies (e.g., automobiles, drones, robotics, etc.) become more sophisticated, they are increasingly capable of performing tasks that were once thought to require human workers. AI powered automation of tasks raises concerns relating to mass job displacement (typically affecting the most vulnerable), and the potential for widespread unemployment, which could impact labor markets and social welfare, potentially leading to business upheaval, industry collapse, economic disruption, and social unrest. AI also has the potential to amplify and exacerbate existing power imbalances, economic disparities, and social inequalities.
- Autonomous weapons:** AI controlled weapons systems could make decisions about when and who to target, or potentially make life-and-death decisions (and kill indiscriminately) without human intervention, raising concerns about ethical implications and potential unintended consequences. Indeed, the development and proliferation of autonomous weapons (including WMDs), and the competition among nations to deploy weapons with advanced AI capabilities, raises fears of a new arms race and the increased risk of a nuclear

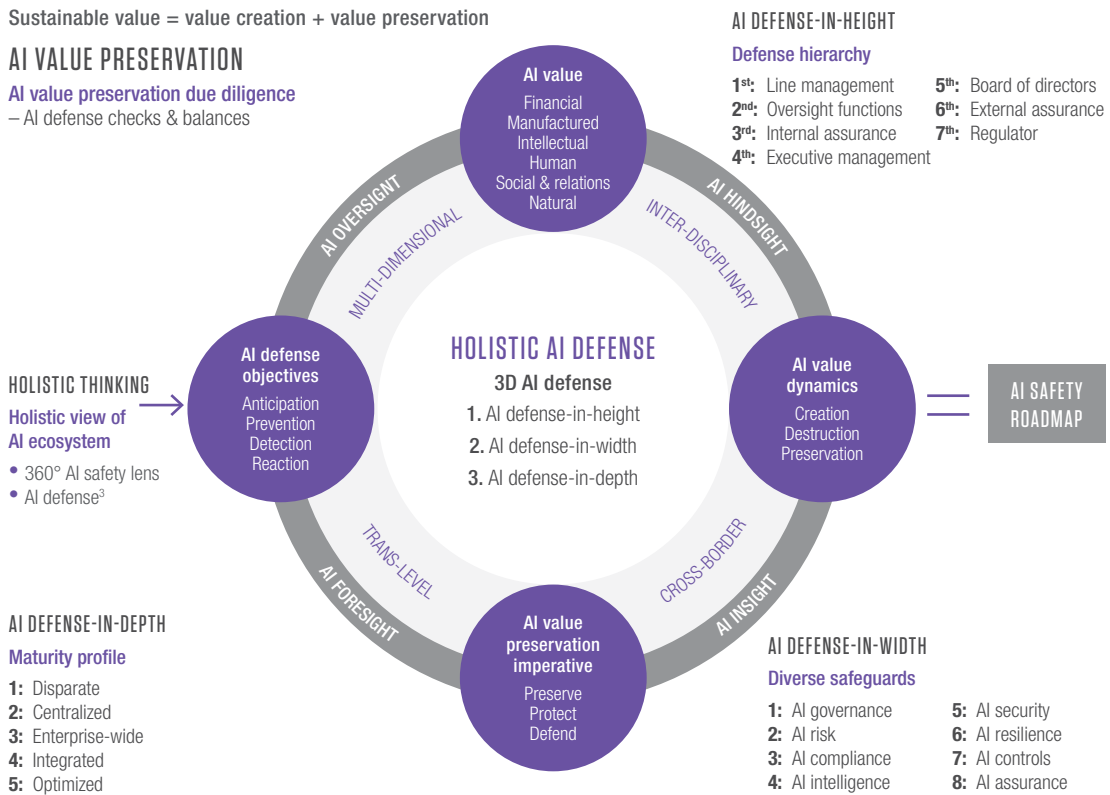
war. This potential for misuse and possible unintended catastrophic consequences could ultimately pose a threat to international security, global safety, and, ultimately, humanity itself.

- Superintelligence and the singularity:** the ultimate threat potentially posed by the AI singularity or superintelligence is a complex and uncertain issue that may (or may not) still be on the distant horizon. The potential for AI to surpass human control and pose existential threats to humanity cannot, and should not, be dismissed, and it is imperative that the appropriate safeguards and controls are in place to address this existential risk. The very possibility that AI could play a role in human extinction should at a minimum raise philosophical questions about our ongoing relationship with AI technology and our required duty of care. Existential threats cannot be ignored and addressing them cannot be deferred or postponed.

2. AI SAFETY DUE DILIGENCE

AI safety includes delivering trustworthy, responsible, and ethical AI systems. AI safety, therefore, involves ensuring that due diligence is rigorously applied throughout the AI safety process. This due diligence consists of adopting a comprehensive and systematic approach, and requires considerable preparation, vigilance, and perseverance on an ongoing basis. Given the nature of the AI safety challenge and the dangers associated with AI risks, threats, and hazards, effective AI safety will require robust protocols, sometimes referred to as the buttons, belts, and braces or the full metal jacket approach. To help ensure confidence and trust in our AI systems, appropriate checks and balances need to be in place and all necessary safeguards and guardrails need to be operating effectively on an ongoing basis.

Figure 2: AI safety roadmap



2.1 AI safety and holistic thinking

AI safety is concerned with defending against the implications of AI dangers, which can result from AI risks, threats, and hazards, all of which are also continuously evolving, adapting, and mutating. Effectively addressing the AI safety challenge demands a holistic mindset to fully understand and appreciate the complicated challenges and complex dynamics posed by developments in AI technology [Google Deepmind (2024)]. In this context holistic thinking involves developing a Gestalt-like understanding of how AI-related issues are intertwined, interconnected, and interdependent. Holistic thinking involves developing a comprehensive view and can incorporate a consolidation of different forms of integrated thinking (e.g., strategic thinking, systems thinking, design thinking, etc.). When addressing AI safety challenges, holistic thinking can help to minimize the disparate flaws, deficiencies, and weaknesses that are likely to be a common feature of future AI safety failures.

2.2 The AI safety ecosystem

Holistic thinking is essential in the development of a comprehensive view of the entire AI ecosystem to gain a better understanding of the AI environment in its totality [WEF (2024)]. The AI landscape of 2024 is sophisticated, dynamic, and constantly evolving in its many different forms. A holistic mindset is necessary to fully appreciate the complicated and complex challenges posed by the rapid developments in the AI technology space.

2.3 A comprehensive approach to AI safety

Naturally, a comprehensive approach to AI safety requires a holistic view to develop the capability to design an extensive AI safety program [Lyons (2024c)]. Holistic AI safety involves viewing circumstances through a 360° AI safety lens and considering, assessing, and evaluating AI safety matters from multiple angles (e.g., outlooks, perspectives, and points of view). The adoption of a comprehensive approach to AI safety can help reduce blind spots and eliminate any cognitive biases that could later result in being rendered vulnerable to the risks posed by AI. Such an approach is essential to AI safety, and it is important that all stakeholder groups satisfy themselves that their organizations are taking all the necessary and appropriate measures.

3. EXTENSIVE AI VISIBILITY

Holistic thinking also requires extensive visibility to effectively monitor events and gain a thorough understanding of the AI challenge in its entirety. Ironically, it also requires the ability to be able to utilize the full capability of AI technology in this regard.

3.1 AI lines of sight:

Harnessing AI's full potential in the following areas can help improve decision making, which could prove to be indispensable going forward and help eliminate AI blindsight [Dailey (2018)].

- **AI hindsight:** AI technology can be harnessed to effectively learn from the experiences of the past to help identify the reasons behind previous successes and failures in any given sector or field.
- **AI insight:** AI technology can be used to help understand, interpret, and derive valuable knowledge from analyzing available data to help enhance decision making. This can include identifying emerging trends (i.e., signals, patterns, and correlations).
- **AI foresight:** AI technology can be used to help to forecast, anticipate, or predict future trends, which can help with forward planning and preparing for all possible future developments, occurrences, and scenarios.
- **AI oversight:** AI technology can be harnessed to help with overseeing and supervising ongoing practices and activities to help monitor performance and ensure conformance with policies, standards, and guidelines.

4. A BIG PICTURE REALITY

Holistic thinking involves ensuring that the implications of AI safety issues are considered from multiple vantage points. A big picture outlook can facilitate viewing AI safety from all directions and is required to facilitate inclusive collaboration, cooperation, and coordination among stakeholder groups. A comprehensive architectural framework is, therefore, essential [Chen et al. (2024)]. It is especially important in terms of fully understanding the potential for different types of consequences (e.g., intended and unintended consequences), the potential cascade of consequences, and the precise nature of any possible contagion.

4.1 Diverse perspectives

The development of an inclusive scope is essential and issues should be considered from the following diverse perspectives:

- **Interdisciplinary:** issues should be considered from an interdisciplinary perspective (i.e., science, law, ethics, sociology, psychology, education, healthcare, etc.) to help ensure the necessary diversity of expertise.
- **Cross-border:** issues should be considered from a cross-border perspective (i.e., local, national, international, global, etc.) in order to help identify anomalies and ensure consistency across international boundaries and jurisdictions.
- **Trans-level:** issues should be considered from a trans-level perspective (i.e., macro, meso, micro, etc.) in order to help ensure greater worldwide alignment of all AI activities, including on strategic, tactical, and operational issues.
- **Multi-dimensional:** issues should be considered from a multi-dimensional perspective (i.e., time, space, matter, consciousness, etc.) to help develop a truly holistic appreciation and understanding of evolving AI and cyberspace realities (i.e., digital reality, augmented reality, virtual reality, etc.).

5. STAKEHOLDER AI VALUE

Stakeholders refer to all those with a vested interest in the activities of a particular organization or group. Stakeholder groups can generally include governments, civil society, private sector, scientific community, and others. In business, stakeholders can include shareholders, board members, management, employees, customers, clients, business partners, regulators, and the public. AI stakeholders can also include users, developers, researchers, policymakers, and investors. All stakeholder groups have a duty of care to ensure that the best interests of their own stakeholders are being taken into consideration [Sharma (2024)].

5.1 AI value

The value utility associated with AI is ultimately determined by its stakeholders. In order to address AI safety, it is important to first gain an understanding of the precise nature of AI value and be able to view AI safety through a value-centric lens. This challenge can begin with an understanding and appreciation

of the evolving concept of AI value (and value drivers) and then proceed to how best manage this notion of AI value once it has been clearly identified. Value utility is increasingly being viewed in the context of society, the economy, and the environment (also referred to as the triple bottom line of people, profit, and planet). In the past, the promise of value was perhaps often associated with price, however, there is now a requirement to also consider value propositions in terms of financial and non-financial value, tangible and intangible value, intrinsic and extrinsic value, and quantitative and qualitative value. As a result, in a multi-stakeholder environment the concept of value is increasingly being viewed in the context of a multi-capital approach.

In the “multi-capital model”, stakeholder AI value can be viewed in terms of the six forms of capitals that all organizations depend on for their success [IIRC (2021)].

- **Financial capital:** financial value is viewed in terms of the value associated with financial capital and primarily relates to financial matters.
- **Manufactured capital:** manufactured value is viewed in terms of the value associated with manufactured capital and primarily relates to physical goods and services.
- **Intellectual capital:** intellectual value is viewed in terms of the value associated with intellectual capital and primarily relates to knowledge-based intangibles.
- **Human capital:** human value is viewed in terms of the value associated with human capital and primarily relates to the value of people.
- **Social and relationship capital:** social and relationship value is viewed in terms of the value associated with social and relationship capital and primarily relates to information sharing networks.
- **Natural capital:** natural value is viewed in terms of the value associated with natural capital and primarily relates to environmental resources.

In practice, the process of increasing any one of these capitals can result in decreasing one or more of the other capitals, resulting in a value trade-off. Each organization must, therefore, identify its own priority stakeholders and determine the type of value that they intend to deliver on behalf of these stakeholders.

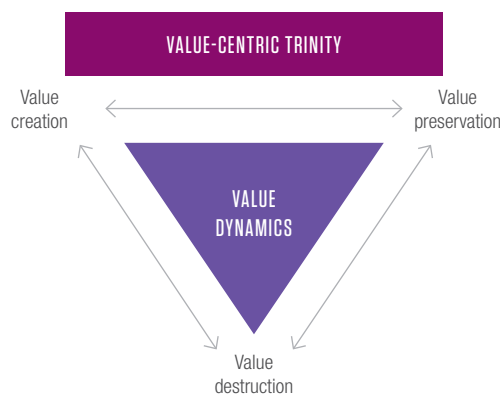
5.2 AI value dynamics

In nature, the primary forces that underpin universal development are represented by creation, preservation, and destruction, which can be evidenced at both the micro (atomic) and the macro (cosmic) level. AI value management involves arriving at a healthy balance between these universal forces as they apply to AI value. Sound AI value management should, therefore, focus on appreciating, understanding, and managing the dynamics of these universal forces. The value-centric trinity acknowledges the existence of these primary universal forces in the context of the management of value and captures the dynamics of their relationship. In this context, these universal forces are represented by AI value creation, AI value preservation, and AI value destruction, which are in continuous interaction with one another [Lyons (2022)].

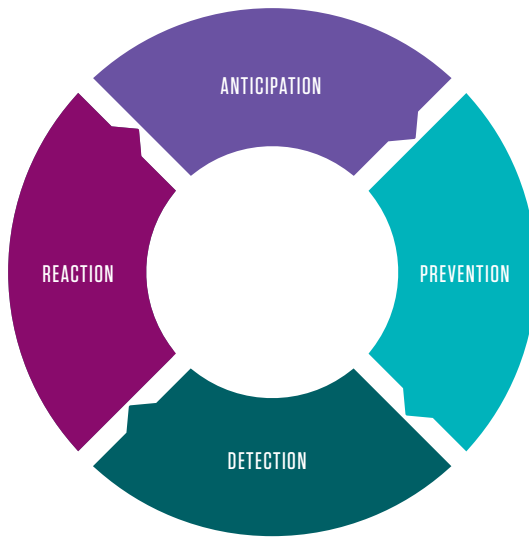
- AI value creation:** value creation is typically associated with enhancing value, increasing value, and generating value. Examples of how AI can create value for its stakeholders include efficiency and productivity, enhanced decision making, personalization, cost reduction, innovation, risk management, and scalability. Typically, business organizations have explicitly addressed the value creation imperative at a strategic level through their company culture, purpose, vision, and business strategy. AI value creation is primarily concerned with exploiting the upside and delivering rewards to its stakeholders. AI value creation is associated with all the creative and exciting activities within the organization. Consequently, it is considered a top priority for most organizations, and it tends to be at the front of people’s minds when it comes to decision making. Those charged with value creation responsibilities generally possess considerable authority, status, and influence within their organizations.

- AI value preservation:** value preservation is associated with safeguarding and future-proofing AI value. Examples of how AI can preserve value for its stakeholders include data security and privacy, bias mitigation, transparency and explainability, continuous monitoring and maintenance, ethical AI practices, regulatory compliance, and stakeholder engagement. Value preservation is concerned with mitigating the downside and is, therefore, often seen as a necessary evil with certain negative connotations. Consequently, value preservation tends to be considered less of a priority and often tends to be considered as an afterthought rather than being part of the initial decision making process.
- AI value destruction:** value destruction is associated with destroying and decreasing stakeholder AI value. Examples of how AI can destroy stakeholder value include matters such as environmental sustainability and destruction, misuse and abuse, privacy, criminality and discrimination, job displacement and social impact, autonomous weapons, and superintelligence and the singularity. These issues have already been addressed in more detail above. AI value destruction can occur at strategic, tactical, and operational levels and it is often difficult to predict the potential knock-on consequences and impact of an initial operational issue. Indeed, it is possible for a seemingly minor incident to cascade into a major crisis if left unchecked. Generally speaking, value destruction is to be avoided and/or minimized, however there may be occasions whereby a certain level of value destruction is regarded as acceptable. As with evolution in nature, sometimes in order to create space for additional AI value creation a certain level of value destruction may be required. In such circumstances, this value destruction is considered to be necessary and is viewed as being intentional and deliberate.

Figure 3: Value dynamics



All types of AI value will be subject to these value dynamics both individually and collectively. Consequently, there needs to be an appreciation of the complexities of these dynamics within the value-centric trinity. In reality, these ongoing interactions are in a constant state of flux and from time to time can require delicate trade-offs between the different forms of AI value. For example, an increase in AI financial capital may be offset by a corresponding decrease in AI natural capital.

Figure 4: Defense cycle

5.3 AI value preservation imperative

Logically, the delivery of sustainable AI value over the short, medium, and long term requires a healthy balance between the focus on value creation and the focus on value preservation in all decision making at strategic, tactical, and operational levels. In nature, in business, and in AI, once something of value has been created it then needs to be safeguarded to survive and to be considered sustainable.

Value preservation is, therefore, primarily concerned with the avoidance of value destruction; however, its broader purpose is to also support continued value creation, which is necessary for long-term survival and sustainability. It is primarily concerned with safeguarding and futureproofing stakeholder AI value and needs to be regarded as a necessary and positive investment in a sustainable future. Value creation and value preservation, therefore, should be addressed in tandem as they go hand-in-hand and could be said to represent two sides of the same coin.

The AI value preservation imperative refers to a duty of care, being the social, moral, and ethical obligation to preserve, protect, and defend stakeholder AI value from value destruction. AI value preservation is focused on defending against hazard events and it is concerned with mitigating risks, protecting against threats, and minimizing vulnerability to hazard events [USDHS (2024)]. Ultimately, it is concerned with defending AI value against all forms of value destruction, including value erosion, reduction, and depletion.

5.4 AI defense objectives

AI defense is synonymous with AI safety and AI value preservation. An iterative defense cycle addresses the key drivers that should be present in all AI defense related activities.

“Unifying defense objectives” represent the necessary drivers of any AI defense mission and consist of the following:

- **Anticipation:** refers to the timely identification and assessment of existing risks, threats, and vulnerabilities, as well as the prediction of future risks, threats, and vulnerabilities.
- **Prevention:** refers to taking sufficient measures to shield against anticipated risks, threats, and vulnerabilities.
- **Detection:** refers to the identification of activity types (e.g., exceptions, deviations, and anomalies) that indicate a breach of defense protocol.
- **Reaction:** refers to the timely response to a particular event or series of events to both mitigate the current situation and to take further corrective action in relation to identified deficiencies.

These drivers represent the cornerstones of an AI defense cycle and represent four essential elements in any AI defense program.

6. HOLISTIC AI DEFENSE

A holistic approach to an AI defense program requires a comprehensive three-dimensional framework, also referred to as 3D AI defense or AI defense cubed (AI defense³).

6.1 AI defense-in-height

AI defense-in-height involves value preservation via an oversight hierarchy that incorporates both internal and external stakeholder lines of defense. Internal lines of defense refer to the hierarchy present along the vertical axis, which incorporates the top-down delegation of authority and assignment of responsibility, with the bottom-up provision of assurance and enforcement of accountability. Oversight includes the supervision of all AI defense activities from the top of the organization or group (i.e., boardroom) to the bottom of the organization (i.e., front lines). Effective AI safety oversight requires competent and capable leadership at all tiers (i.e., strategic, tactical, and operational) of an organization or group.

A complete oversight framework should incorporate the traditional “three lines of defense” model with executive management and the board of directors as the all-important fourth and fifth strategic lines of defense as follows:

- **Operational line management:** as the first line of defense, operational line management (i.e., front, middle, and back office) is responsible for overseeing all day-to-day operations and activities of the AI defense program.
- **Tactical oversight functions:** as the second line of defense, tactical oversight functions (i.e., risk management, compliance, security, etc.) are responsible for the oversight of operational line management and for providing subject matter expertise, guidance, and tactical support in relation to AI defense matters.
- **Independent internal assurance:** as the third line of defense, independent internal assurance (i.e., internal audit) is responsible for reviewing the activities of the first and second lines of defense and for providing independent assurance on the effectiveness of the AI defense program.
- **Executive management:** as the fourth line of defense, executive management is responsible for providing AI defense leadership and for providing assurance to the board of directors that the objectives of the AI defense program are being achieved.
- **Board of directors:** as the fifth and last line of defense, the board of directors has overall responsibility for AI defense oversight and is accountable to stakeholders for the program’s strategy and performance.

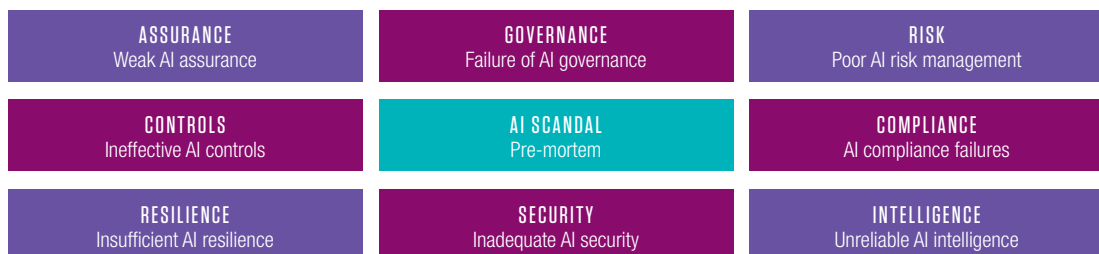
AI safety oversight by external gatekeepers and watchdogs can help to address the separation of power issue that is an inherent flaw present in self-regulation and in voluntary adherence. This can include various sources of external assurance (e.g., validation, certification, ratings, etc.) and

the oversight and supervision by the relevant regulator (i.e., national, international, and global). AI defense-in-height requires transparency and accountability in relation to the competence, capability, and performance of those (individuals and groups) charged with oversight responsibilities. This is critical for establishing and maintaining confidence and trust in the AI safety ecosystem.

6.2 AI defense-in-width

AI defense-in-width involves value preservation through diversity and ensuring that AI challenges are viewed from different perspectives (and through different lenses) to help ensure fairness, minimize cognitive biases, and eliminate potential blind-spots. This requires the sharing of information and exchange of knowledge across the horizontal axis, which includes trans-organizational, interdisciplinary, and cross-functional, collaboration, cooperation, and coordination. Defense-in-width requires an inclusive and integrated approach incorporating a wide spectrum of expertise, experience, and skills within an organization. In particular, it must specifically involve both an individual and a collective focus on the eight critical AI defense components (i.e., AI governance, AI risk, AI compliance, AI intelligence, AI security, AI resilience, AI controls, and AI assurance). Individually, these components can help provide different layers of defense and collectively they can actually fortify and reinforce one another. Each of these eight critical AI defense components are interconnected, intertwined, and interdependent as individually each impacts on, and is impacted by, each of the other components. They represent links in a chain where the chain is only as strong as its weakest link. Individually and collectively, they can provide diverse safeguards and guardrails, but perhaps more importantly they can help to create an essential cross-referencing system of checks and balances to help ensure that AI activities are safe, ethical, and legal.

Figure 5: AI scandal pre-mortem



Conversely, post-mortem investigations into the causes of corporate scandals typically identify flaws, deficiencies, and weaknesses in these eight critical components [Lyons (2016)], whereby their existence in more than one of these critical components can collectively result in exponential collateral damage to stakeholder value. It is, therefore, reasonable to foresee that these same weaknesses are also likely to arise in relation to future AI scandals [Lyons (2024a)].

Prudence and common sense would suggest that it is considered both logical and rational to anticipate the following weaknesses in relation to AI technology and to fully consider their potential for value destruction.

- **Failures in AI governance:** the current lack of a single comprehensive global AI governance framework has already led to inconsistencies and differences in approaches across various jurisdictions and regions [U.N. (2024)]. This is likely to result in potential conflicts between stakeholder groups with different priorities. The lack of a unified approach to AI governance can result in a lack of transparency, responsibility, and accountability, which raises serious concerns about the social, moral, and ethical development and use of AI technologies. The ever-increasing lack of human oversight due to the development of autonomous AI systems simply reinforces these growing concerns.
- **Poor AI risk management:** currently, there appears to be a fragmented global approach to AI risk management. Some suggest that this approach seems to overemphasize a focus on risk detection and reaction and underemphasize a focus on risk anticipation and prevention. It can tend to focus on addressing very specific risks (e.g. bias, privacy, security, etc.) without giving due consideration to the broader systemic implications of AI development and its use [MIT Future Tech (2024)]. Such a narrow focus on AI risks also fails to address the broader societal and economic impacts of AI and overlooks the interconnectedness of AI risks and their potential long-term consequences. Such short-sightedness is potentially very dangerous as it fails to address and keep pace with the potential damage of emerging risks while also failing to prepare for already flagged longer-term risks such as those posed by superintelligence or autonomous weapons systems, among others.
- **AI compliance failures:** AI compliance consists of a patchwork of AI laws, regulations, standards, and guidelines at national and international levels. This lack of harmonization of laws and regulations means that they are not in clear alignment, meaning they can be inconsistent in nature. This makes them both confusing and ineffective, making it difficult for stakeholders to comply with, and for regulators to supervise and enforce, especially across borders [E.U. (2024)]. This lack of clear regulation, as well as a lack of appropriate enforcement mechanisms makes it difficult to hold actors to account for their actions and can encourage non-compliance, violations, and serious misconduct leading to the potential unsafe, unethical, and illegal use of AI technology. The existence of algorithmic bias can result in a lack of fairness and lead to an exacerbation of existing inequality, prejudice, and discrimination. A major concern is that the current voluntary nature of AI compliance and an overreliance on self-regulation is not sufficient to address these potentially systemic issues.
- **Unreliable AI intelligence:** unreliable intelligence can ultimately result in poor decision making in its many forms. Many AI algorithms can be opaque in nature and are often referred to in terms of a “black box”, which hinders the clarity and transparency of the development and deployment of AI systems. Their complexity makes it difficult to interpret or fully comprehend their algorithmic decision making and other outputs [ICO (2020)]. It is, therefore, difficult for stakeholders to understand and mitigate their limitations, potential risks, and the existence of biases. This can further contribute to accountability gaps and make it difficult to hold AI developers and users accountable for their actions. AI development can also lack the necessary stakeholder engagement and public participation, which can mean a lack of the required diversity of thought needed for the necessary alignment with social, moral, and ethical values.
- **Inadequate AI security:** the global approach to AI security also appears to be somewhat disjointed. Data is one of the primary resources of the AI industry and AI systems collect and process vast amounts of data. AI technologies can be vulnerable to cyberattacks, which can compromise assets (including sensitive data), disrupt operations, or even cause physical harm. If AI systems are not properly protected and secured, they could be infiltrated or hacked, resulting in unauthorized access to data, which could be used for malicious purposes such

as data manipulation, identity theft, or fraud. This raises concerns about data breaches, data security, and personal privacy [NCSC and CISA (2023)]. Indeed, AI powered malware could help malicious actors evade existing cyber defenses, thereby enabling them to inflict significant destruction to supply chains and critical infrastructure (e.g., damage to power grids and disruption of financial systems, etc.).

- **Insufficient AI resilience:** the global approach to AI resilience is naturally impacted by the chaotic approach to some of the other areas noted above. Where AI systems are vulnerable to cyberattacks, this can allow hackers to disrupt operations, leading to possible unforeseen circumstances that are difficult, if not impossible, to prepare for. This could impact the reliability and robustness of the AI system, its ability to perform as intended in real-world conditions, and to withstand, rebound, or recover from a shock, disturbance or disruption. AI systems can, of course, also make errors, incorrect diagnoses, faulty predictions, or other mistakes. Where an AI system malfunctions or fails for whatever reason, this can lead to unintended consequences or safety hazards that could negatively impact on individuals, society, and the environment [CSA (2024)]. This may be of particular concern in terms of the preparedness of critical domains such as power, transportation, health, and finance.
- **Ineffective AI controls:** the global approach to AI controls also seems to be somewhat disorganized. Once AI systems are deployed [IBM (2024)], it can be difficult to change them. This can make it difficult to adapt to new circumstances or to correct mistakes. There are, therefore, some concerns that an overemphasis on automated technical controls (such as bias detection and mitigation etc.) and not enough attention given to the importance of human control can create a false sense of security and mask the need for human control mechanisms. As AI systems become more sophisticated, there is a real risk that humans will lose control over AI, leading to situations where AI may make decisions that have unintended consequences that can significantly impact on individuals' lives with potentially harmful consequences. Increasing the autonomy of AI systems without the appropriate safeguards and controls in place raises valid concerns about issues such as ethics, responsibility, accountability, and potential misuse.

- **Weak AI assurance:** there is currently no single, universally accepted framework or methodology for AI assurance. Different organizations and countries have varying approaches, leading to potential inconsistencies. The opaque nature and increasing complexity of AI can make it difficult to competently assess AI systems, creating gaps in assurance practices, and thus hindering the provision of comprehensive assurance [Batarseh and Freeman (2022)]. The expertise required for effective AI assurance is often a scarce commodity and may be unevenly distributed, which, in turn, can create accessibility challenges for disadvantaged areas and groups. The lack of transparency, ethical concerns, and the lack of comprehensive AI assurance can lead to an erosion of public trust and confidence in AI technologies, which can hinder its adoption and potentially create resistance to its potential benefits. Given all of the above, the provision of AI assurance can be a potential minefield for assurance providers.

6.3 AI defense-in-depth

AI defense-in-depth involves value preservation through developments in maturity and formality that reflect the general attitude to AI safety in terms of culture, mindset, and DNA. Robust AI defense-in-depth requires appropriate levels of maturity across the entire organization, particularly across all the critical AI defense components (both individually and collectively). AI defense-in-depth refers to the level of maturity present throughout the front to back axis, which reflects the insights, knowledge, and wisdom present within the organization or group. A focus on defense-in-depth helps to ensure that defense-in-height and defense-in-width measures are not just theoretical in nature, simply window dressing, or merely AI defense theatre. Defense maturity can be ascertained by the extent to which the current AI defense approach has developed by chance or by design. The maturity profile can indicate the strength of AI defense in practice.

Typically, the "maturity profile" indicates the level of maturity and formality in place and can be plotted on a safety or defense spectrum [Dalrymple et al. (2024)], or simply classified in terms of the different phases of a standard maturity model [Lyons (2016)] as follows:

- **Disparate phase:** AI defense activities operate in a fragmented approach, where processes are developed on an ad-hoc and inconsistent basis. This can result in matters being addressed in an unsystematic,

unstructured, and reactive manner that can lead to crisis mode operations and continuous firefighting on a day-to-day basis.

- **Centralized phase:** AI defense activities have centralized competence centers of dedicated individuals with specialized skills and expertise. As a defined professional discipline, basic policies, procedures, and practices are established so that they can be repeated.
- **Enterprise-wide phase:** AI defense activities have agreed principles and processes that operate throughout the organization or group so that common practices are adopted on an enterprise-wide basis in a systematic and structured manner. Defined objectives and methodologies are standardized and documented.
- **Integrated phase:** AI defense activities utilize technology for end-to-end vertical and horizontal integration (i.e., people, processes, and systems). This enables effective management and the meaningful reporting of essential measurement metrics relating to performance and productivity. Processes are measured and controlled.
- **Optimized phase:** AI defense activities focus on deliberate process upgrading and optimization of resources. This facilitates workforce empowerment through enhanced performance and constant efforts at continuous improvement, accelerated learning, and pioneering innovation.

The AI defense spectrum can vary widely in terms of maturity, capability, and competency. For example, they can range from implicit, informal, undocumented, and unstructured programs on the one hand, to explicit, formal, documented, and structured programs on the other hand, and everything else in between. This can include the existence (or non-existence) of a formally documented and approved AI defense charter (including purpose, vision, mission statement, strategy, framework, plan, policies, procedures, etc.). Immature programs often operate in a rather chaotic or disorganized manner, as they often lack a sense of a unifying structure and a systematic approach. The degree to which the program is explicit, formal, documented, and structured represents a clear indication of the organization or group's focus on its AI defense obligation to minimize AI value destruction.

6.4 AI DEFENSE-IN-UNITY: UNIFIED DEFENSE

Ultimately, holistic AI defense involves unifying and uniting all three dimensions within a single framework so that all AI defense activities are strategically aligned, tactically integrated, and operating in unison towards common AI defense objectives. Not surprisingly, when operating together defense-in-height, defense-in-width, and defense-in-depth can provide an organization or group with a higher grade of defense.

Holistic AI defense must be regarded as being dynamic in nature and will require continuous learning, constant improvement, and ongoing refinement. This means utilizing hindsight, insight, and foresight on a permanent basis. Logically, holistic AI defense will improve over time as the defense insights, knowledge, and wisdom also improve over time. Wisdom in AI defense decision making combines the knowledge acquired through past experiences with an understanding of the present environment, and an expectation of future developments.

7. ROBUST AI DEFENSE AND THE AI COMPLEXITY CHALLENGE

It may well be that there are limits to the level of AI complexity that humans can effectively manage and that at some point the level of complexity arising out of technological development will simply become too complex for humans to manage. In the past, the concept of holistic AI defense may perhaps have been considered too difficult and complicated for certain organizations to address. Indeed, it could now be argued that the advancements in AI technology have actually made this challenge even more complex. Ironically, these same advancements in AI technology that rightly raise concerns, also have the potential to make this challenge more manageable, provided this is addressed in a prudent and conscientious manner [Lyons (2024b)].

7.1 The paradox of AI

The paradox of AI is that eventually only AI technology will have the capability to manage the complexity of AI technology. Ironically, it seems increasingly likely that it is only through sophisticated AI technology that humans can ever hope to effectively manage the increasing complexities of the digital world. For this to occur in as ethical, safe, and secure a manner as possible it will, however, require enhanced levels of AI safety due diligence. Such an approach can help contribute to a more peaceful and secure world, by creating a more trustworthy, responsible, and beneficial AI ecosystem for all.

7.2 Leveraging AI technology

AI technology can now be leveraged to enhance the management of AI defense by supporting, supplementing, and augmenting human capabilities in this space. Holistic AI defense is now a realistic expectation because of AI's growing superpowers in an increasing number of disciplines, in which its capabilities have already surpassed that of humans. Though still in its infancy, the use of AI to supplement human capabilities in this field is already occurring in many of these areas, particularly in the cyber defense space (e.g., cyber intelligence, cybersecurity, cyber resilience, etc.). This potential comes with notable health warnings. A holistic approach to AI defense is now increasingly possible by employing these evolving AI superpowers, however this too needs to be done in a safe and secure manner. With the necessary safeguards in place, it becomes possible to harness AI's transformative potential and utilize its decision making and problem-solving capabilities to help unlock new opportunities.

7.3 AI defense fortification

The challenge of upgrading our approach to AI defense is, however, now becoming a realistic proposition due to the ongoing utilization of technology with varying levels of AI sophistication to augment and fortify defense related activities as follows:

- **Diligence:** by embedding due diligence into the AI lifecycle (i.e., ideation, design, development, deployment, maintenance, and retirement), organizations can better adhere to best practices and help ensure fairness, minimize bias, and eliminate discrimination. For example, data is generally considered to be the lifeblood of AI and the success of its performance is very much dependent on the quality, quantity, and provenance of data used throughout its lifecycle. Data robustness can be improved by incorporating the critical AI defense components into the data management framework (e.g., data governance, data risk, data compliance, data intelligence, data security, data resilience, data controls, and data assurance).
- **Automation:** advanced technology (including the use of AI bots) can be used to automate the activities of these critical AI defense components and to help to ensure that these activities are autonomously operating on a continuous basis and providing real-time information. Ongoing activities such as verification, validation, and testing can benefit from automation and help to increase confidence and trust in defense processes (e.g., automated auditing, continuous auditing, real-time auditing, etc.).
- **Specialization:** the use of specially focused narrow AI (e.g., algorithms, analytics, models, platforms, etc.) can be used to perform specific AI defense activities from cradle to grave. This can involve narrow technical solutions and can include processes such as issue identification, assessment, remediation, monitoring, and reporting (e.g., risk identification, risk assessment, risk response, risk monitoring, risk reporting, etc.).
- **Foresight:** forward looking and future focused technologies can be used as forecasting instruments and tools to help support the anticipation of future issues. Foresight enables the implementation of proactive measures in advance. These technologies can involve the use of predictive analytics, sensitivity analysis, scenario modeling, and scenario simulations (e.g., resiliency analysis, predictive maintenance, crisis modeling, scenario testing, etc.).
- **Interconnectivity:** AI technology can be used to help better understand symbiotic relationships and appreciate the correlations, dependencies, and interconnectivity of activities. This can involve the extrapolation of first, second, and third order consequences to outline any possible cascades of contagion. This can help to create, protect, and maintain a big picture perspective (e.g., relational mapping, interconnectivity linking, and consequence projections).
- **Speed:** the use of technology can help to contain potentially volatile situations from quickly escalating by helping to accelerate reactions and speed up response times. The timely detection of unusual, unexpected, abnormal, or suspicious activity can be critical. This can help ensure that an individual incident does not escalate to an emergency, to a crisis, to a disaster, and on to a catastrophe (e.g., real-time alerts, early warning mechanisms, various response triggers, etc.).
- **Learning:** the use of self-learning technology offers the potential of continuous learning in real-time based on learning from ongoing behaviors, subtle patterns, and performance metrics. Adaptive learning capabilities can help defense activities to evolve and develop on a day-to-day basis, thereby helping to amplify defense processes, enhance defense capabilities, and improve the overall defense posture (e.g., adaptive authentication, adaptive recovery, adaptive controls, etc.).

- **Vigilance:** technology can be used to help improve vigilance in terms of the current environment. Real-time vigilance can help to ensure early intervention and adherence to frameworks, codes, best practices, and standards, thereby helping to minimize the occurrence of negative events. The quality of corporate health can be monitored using diagnostics to indicate potential compromises and violations (e.g., anomalies, deviations, system failures, etc.), which can help to quickly identify new exposures, vulnerabilities, and operational gaps (e.g., scanning technology, benchmarking tools, exception reporting, etc.).
- **Decisions:** AI technology can be used to enhance, augment, and support decision making through education, training, and awareness, thereby helping improve options and choices. AI driven personalization based on professional and personal preferences can provide tailored content and recommendations through customized updates, guidance, and assistance. AI can help provide the individual with the transparency required to arrive at more informed, ethical, and risk-weighted decisions (e.g., explainable AI (XAI), user-friendly interfaces, virtual assistants, etc.).
- **Collaboration:** AI technology can help facilitate stakeholder interactions, collaboration, cooperation, and coordination through group communication interfaces. It can facilitate group brainstorming in addition to the constant sharing of ideas and insights, and the ongoing exchange of information, intelligence, and knowledge as part of the collaboration process (e.g., chat platforms, chatrooms, chatbots, etc.).

8. CONCLUSION

This article presents a high-level outline of a possible AI safety roadmap to help ensure the development of trustworthy, responsible, and ethical AI around the world. Global AI safety is critical to defend against the potential downside of AI (from routine to existential risks) and needs to be prioritized accordingly. Our global leaders have a duty of care to safeguard against the potential damage of this impending AI value destruction and that will require a much higher, more robust, and more mature level of AI safety due diligence than is currently on display. Dynamic developments in AI mean that the normal order of things no longer applies and that going forward effective AI safety will require superior levels of guardianship, stewardship, and leadership.

In practice, effective AI safety measures require the highest preemptive capabilities to be in place because it is the reaction times to potentially devastating events that will determine the magnitude of the initial impact and the subsequent collateral damage. AI safety requires a harmonization of global, international, and national frameworks, regulations, and practices to help ensure consistent implementation and the avoidance of fragmentation. This means greater coordination, knowledge exchange, and information sharing to help ensure a robust and equitable global AI safety environment.

REFERENCES

- Batarseh, F. A., and L. Freeman, 2022, "AI assurance: towards trustworthy, explainable, safe, and ethical AI," Academic Press, <https://tinyurl.com/3jxjjeed>
- Chen, C., Z. Liu, W. Jiang, S. Q. Goh, and K-Y. Lam, 2024, "Trustworthy, responsible, and safe AI: a comprehensive architectural framework for AI safety with challenges and mitigations," arXiv, <https://tinyurl.com/y773yv6j>
- CSA, 2024, "AI resilience: a revolutionary benchmarking model for AI safety," Cloud Security Alliance, May, <https://tinyurl.com/yfijxyxy>
- Dailey, P., 2018, "On governance: balancing directors' hindsight, insight, and foresight for board composition and effectiveness," The Conference Board, May, <https://tinyurl.com/mtwwmbxx>
- Dalrymple, D., et al., 2024, "Towards guaranteed safe AI: a framework for ensuring robust and reliable AI systems," arXiv, May, <https://tinyurl.com/2s453wat>
- E.U., 2024, "The Artificial Intelligence Act (Regulation (EU) 2024/1689)," European Union, August, <https://tinyurl.com/mnv48dew>
- FLI, 2023, "Pause giant AI experiments: an open letter," Future of Life Institute, March, <https://tinyurl.com/5n8uj3s6>
- Google Deepmind, 2024, "Holistic safety and responsibility evaluations of advanced AI models," April, <https://tinyurl.com/4nax2zub>
- IBM, 2024, "Generative AI controls framework: safe, secure, and compliant AI adoption approach," whitepaper, June, <https://tinyurl.com/3zneb5ts>
- ICO, 2020, "Explaining decisions made with AI," Information Commissioner's Office and The Alan Turing Institute, May, <https://tinyurl.com/5u6ewrns>
- IIRC, 2021, "International <IR> Framework," International Integrated Reporting Council, January, <https://tinyurl.com/y5pwe7jr>
- Lyons, S., 2024a, "Pre-mortem of an A.I. scandal(s): anticipation of future hazards," LinkedIn, February, <https://tinyurl.com/2s43u5t2>
- Lyons, S., 2024b, "A.I. value preservation and the paradox of A.I.," LinkedIn, May, <https://tinyurl.com/bddywjds>
- Lyons, S., 2024c, "AI safety roadmap: the AI value preservation imperative," LinkedIn, June, <https://tinyurl.com/5kwuxhy7>
- Lyons, S., 2022, "Value, value proposition, and value management," LinkedIn, September, <https://tinyurl.com/mryaek7b>
- Lyons, S., 2016, Corporate defense and the value preservation imperative: bulletproof your corporate defense program, CRC Press, Taylor & Francis Group, <https://tinyurl.com/yfsxjz9n>
- Mazzucato, M., 2024, "The ugly truth behind ChatGPT: AI is guzzling resources at planet-eating rates," The Guardian, May, <https://tinyurl.com/ztwxy7u2>
- MIT Future Tech, 2024, "AI risk repository: a comprehensive database of risks from AI systems," August, <https://tinyurl.com/5fnudurb>
- NCSC and CISA, 2023, "Guidelines for secure AI system development," U.K. National Cyber Security Centre and the U.S. Cybersecurity and Infrastructure Security Agency, November, <https://tinyurl.com/3wyzdrt>
- NIST, 2024, "NIST trustworthy and responsible AI NIST-AI-600-1 – Artificial intelligence risk management framework," National Institute of Standards and Technology, July, <https://tinyurl.com/yssrdfw6>
- OSTP, 2022, "Blue print for an AI bill of rights: making automated systems work for the American people," The White House Office of Science and Technology Policy, October, <https://tinyurl.com/yeyajm7z>
- Sharma, S., 2024, "Benefits or concerns of AI: A multistakeholder responsibility," Futures 157, March, <https://tinyurl.com/3c7ju8ry>
- U.N., 2024, "Governing AI for humanity," United Nations AI advisory body, September, <https://tinyurl.com/37r4t54x>
- USDHS, 2024, "Artificial intelligence roadmap 2024," United States Department of Homeland Security, March, <https://tinyurl.com/5dkmuj3f>
- WEF, 2024, "Responsible AI playbook for investors," Whitepaper, World Economic Forum, June, <https://tinyurl.com/ykb9e8ue>

GENERATIVE AI TECHNOLOGY BLUEPRINT: ARCHITECTING THE FUTURE OF AI-INFUSED SOLUTIONS

CHARLOTTE BYRNE | Managing Principal, Capco

THOMAS HILL | Principal Consultant, Capco

ABSTRACT

The generative AI (GenAI) landscape is evolving rapidly – and transforming how organizations approach and embrace technology and innovation. As businesses seek to harness the power of GenAI, it is crucial they establish a robust technology blueprint that guides the development, deployment, and management of AI-driven solutions. We explore the essential elements of a GenAI technology blueprint, covering the importance of flexible architectures, ethical considerations, and seamless integration with existing systems.

1. WHY A GENERATIVE AI TECHNOLOGY BLUEPRINT MATTERS

To effectively develop a GenAI technology blueprint, it is essential to recognize that GenAI is not the only factor shaping the future of technology – GenAI's synergy with the broader tech stack (including other artificial intelligence and machine learning tools), as well as the strength of an organization's data foundations, the robustness of past integrations, and the scope of cloud computing capabilities, will all have a profound impact.

A well-defined GenAI technology blueprint will serve as an invaluable roadmap, providing a structured approach to designing and implementing GenAI solutions that align with business objectives, while also addressing the unique challenges posed by GenAI.

By establishing clear architectural principles, governance frameworks, and integration strategies in advance, organizations can ensure the scalability, maintainability, and ethical deployment of GenAI solutions.

2. UNIQUE CHALLENGES POSED BY GENERATIVE AI

GenAI presents a set of distinct challenges that organizations must navigate to ensure successful adoption and deployment, some of which we highlight below.

Data quality and bias. GenAI models rely heavily on the quality and diversity of training data. Ensuring that the data used for training is representative, unbiased, and ethically sourced is a significant challenge. Biased data can lead to discriminatory or unfair outcomes, perpetuating societal biases in AI-generated content.

Intellectual property and content ownership. GenAI models have the ability to generate novel content, such as text, images, and audio. Determining the ownership and intellectual property rights associated with AI-generated content can be complex. Organizations must establish clear guidelines and legal frameworks to address issues related to content ownership, attribution, and licensing.

Explainability and interpretability. GenAI models, particularly deep learning-based models, can be highly complex and opaque. Understanding how these models arrive at their outputs and making their decision making processes interpretable is a significant challenge. Ensuring transparency and explainability is crucial for building trust in GenAI systems and meeting regulatory requirements.

Ethical considerations. GenAI raises ethical concerns related to privacy, fairness, and responsible use. Organizations must grapple with questions such as data privacy, consent, and the potential misuse of GenAI technologies for malicious purposes. Developing ethical frameworks and guidelines is essential to ensure the responsible deployment of GenAI solutions.

Integration with legacy systems. Integrating GenAI solutions with existing legacy systems can be challenging. Organizations must navigate compatibility issues, data integration challenges, and the need for seamless interoperability between GenAI components and traditional software systems. Overcoming these integration hurdles requires careful planning and robust integration strategies.

Talent and skills gap. The rapid advancement of GenAI technologies has created a talent and skills gap. Organizations face the challenge of acquiring and retaining employees with expertise in GenAI techniques, such as deep learning, natural language processing, and computer vision. Building internal capabilities and upskilling the existing workforce are crucial for successful GenAI adoption.

3. KEY COMPONENTS OF A GENERATIVE AI TECHNOLOGY BLUEPRINT

3.1 GenAI application architecture

A GenAI technology blueprint should outline a flexible and scalable application architecture designed to leverage the capabilities of generative models. The architecture should facilitate the creation of new, unique content using enterprise data and integrate seamlessly with current systems for diverse applications.

The following key components should be considered when establishing an application architecture:

- **Experience layer:** this layer encompasses various user interfaces, such as chatbots, contact center portals, web applications, and API playgrounds, enabling seamless interaction with GenAI solutions.

“

Capco realized additional efficiencies of up to 50% in certain tasks where individuals had the right training.

”

- **API management:** robust API management is crucial for facilitating integration between GenAI applications and external systems, ensuring secure and efficient data exchange.
- **GenAI platform:** the GenAI platform serves as the core of the architecture, providing orchestration and model management capabilities. It includes components such as prompt libraries, GenAI models (custom, open-source, and closed-source), and MLOps platforms for model training and deployment.
- **Data storage:** efficient data storage mechanisms, such as knowledge graphs, relational databases, data lakes, and vector databases, are essential for storing and retrieving relevant data for GenAI models.
- **Observability and monitoring:** comprehensive observability and monitoring tools are necessary to track the performance, usage, and outcomes of GenAI solutions, enabling continuous improvement and auditing.

3.2 Types of GenAI models and adaptation strategies

The technology blueprint should consider the various types of GenAI models available and provide guidance on adapting them to specific use cases. The blueprint should cover the following aspects:

- **Model catalog:** maintaining a comprehensive model catalog is crucial for managing and updating information about existing GenAI models, as well as integrating new models as they become available. The catalog should include details such as model types, use cases, performance benchmarks, architectures, and data requirements.

- **Model customization:** the blueprint should outline strategies for customizing GenAI models to specialize in specific domains or tasks. Techniques such as fine-tuning, adapter tuning, and reinforcement learning from human feedback (RLHF) can be employed to enhance model performance and adapt to specific requirements.
- **Retrieval augmented generation (RAG):** RAG is a powerful technique that combines retrieval mechanisms with generative models to provide more accurate and contextually relevant responses. The blueprint should provide guidance on implementing techniques like RAG, including data retrieval strategies, embedding techniques, and integration with GenAI models.
- **Prompt engineering:** effective prompt engineering is crucial for guiding GenAI models to generate desired outputs. The blueprint should cover best practices for crafting prompts, including techniques such as zero-shot, one-shot, and few-shot learning, as well as chain-of-thought prompting and prompt chaining.

3.3 Solution designs and rationale

The GenAI technology blueprint should provide standardized architectural designs and recommendations on how the architectural patterns have been applied to trending GenAI capabilities or similar sets of requirements. It should include:

- **Data architecture:** the blueprint should outline the key data components and considerations for GenAI solutions, such as data discovery, profiling, sourcing, ownership, quality, metadata, and storage.

- **Technology stack:** the blueprint should recommend a suitable technology stack for implementing GenAI solutions, leveraging tools and services from leading cloud platforms such as Microsoft Azure, Amazon Web Services or Google Cloud Platform.
- **Deployment patterns:** the blueprint should provide guidance on deploying GenAI solutions using various patterns, such as containerization, serverless computing, and edge deployment, based on specific requirements and constraints.

3.4 GenAI LLM ops framework

The GenAI technology blueprint should include a framework for building and optimizing Large Language Model Operations (LLM Ops). The LLM Ops framework should cover the following aspects:

- **Model development:** guidelines for selecting the appropriate foundation models, training datasets, and architectures for GenAI model development.
- **Model deployment:** best practices for deploying GenAI models, including considerations for scalability, performance optimization, and monitoring.
- **Model maintenance:** strategies for maintaining and updating GenAI models, including version control, continuous integration and deployment (CI/CD) pipelines, and performance monitoring.
- **Governance and security:** frameworks for ensuring the ethical use, misuse prevention, and adherence to compliance standards in GenAI model development and deployment.

Figure 1: To embark on the journey of creating a GenAI technology blueprint, organizations should consider the following five steps.

1. ASSESS	2. ENGAGE	3. DEVELOP	4. ESTABLISH	5. INVEST
the current state of their technology landscape and identify areas where GenAI can deliver the most value.	with domain experts, AI practitioners, and business stakeholders to gather requirements and align GenAI initiatives with strategic objectives.	a roadmap that outlines the phased implementation of GenAI solutions, considering architectural principles, governance frameworks, and integration strategies.	partnerships with technology vendors, research institutions, and industry consortia to leverage best practices, access cutting-edge tools, and contribute to the broader GenAI ecosystem.	in talent development and upskilling programs to build the necessary expertise in GenAI technologies and ensure a smooth transition to AI-driven solutions.



4. CONCLUSION: GETTING STARTED WITH A GENERATIVE AI TECHNOLOGY BLUEPRINT

A well-crafted GenAI technology blueprint is a vital tool for organizations seeking to harness the transformative power of generative AI. By prioritizing flexible architectures, ethical considerations, seamless integration and continuous monitoring, organizations can accelerate their GenAI adoption and unlock new opportunities for innovation and growth.

As the GenAI landscape continues to evolve, organizations that invest in robust technology blueprints will be ideally positioned to navigate the challenges and opportunities ahead – effectively leveraging GenAI to drive transformative outcomes and shape the future of their industries.

UNLOCKING AI'S POTENTIAL THROUGH METACOGNITION IN DECISION MAKING

SEAN MCMINN | Director of Center for Educational Innovation, Hong Kong University of Science and Technology

JOON NAK CHOI | Advisor to the MSc in Business Analytics and Adjunct Associate Professor, Hong Kong University of Science and Technology

ABSTRACT

The rapid advancement of generative artificial intelligence (GenAI) tools has significant implications for creativity, decision making, and problem solving across various sectors. While AI offers opportunities to enhance productivity by offloading routine tasks to it, excessive or inappropriate dependence can diminish human cognitive engagement and critical thinking skills. This paper highlights the importance of metacognition, which is the ability to reflect on one's thinking and decision making strategies, in effectively integrating AI into both educational and professional settings. By developing metacognitive awareness and employing strategic approaches, individuals and organizations can assess when and how to use AI effectively. Addressing the AI literacy gap is also crucial as it empowers users to navigate AI-driven environments appropriately and confidently. Ultimately, fostering metacognitive skills ensures that AI serves to enhance, rather than replace, human judgment, creativity, and ethical responsibility in decision making processes. This article introduces key metacognitive strategies for effective AI integration and underscores the necessity of continuous learning and human oversight.

1. INTRODUCTION

The rapid development of generative artificial intelligence (GenAI) tools over the past two years has triggered considerable speculation of its impact on creativity, decision making, and problem solving [Chen et al. (2023), Essel et al. (2024), Hao et al. (2024), Kabashkin et al. (2023)]. The World Economic Forum highlights that AI will be both a major job creator and a disruptor [Di Battista et al. (2023)]. Its Future of Jobs Report [Di Battista et al. (2023)] suggests that GenAI tools have already surpassed humans in crucial technical skills like programming, cybersecurity, and design. While observers initially speculated that GenAI tools would displace workers [Hatzius et al. (2023)], more recent speculation has focused on the possibility that AI will empower workers who know how to leverage it to outcompete peers that do not [Lakhani (2023)].

For this reason, training and recruiting AI-ready talent has become crucial for businesses. Within the next five years, technology training programs focusing on AI and big data are

projected to comprise over 40% of the total in companies surveyed across the U.S., China, Brazil, and Indonesia [Di Battista et al. (2023)]. Correspondingly, employers now expect MBA graduates, for instance, to be proficient in leveraging GenAI tools [Jones and Olson (2024)]. While MBAs and other business graduates will not need to be technical experts in AI, they will nevertheless be expected to understand how to leverage GenAI in conjunction with traditional skills such as managing interpersonal relationships, working collaboratively, and leading teams [Jones and Olson (2024)].

Yet, universities have yet to adequately prepare graduates to properly leverage GenAI. The Digital Education Council (henceforth, DEC) Global AI Student Survey (2024) suggests that post-secondary school students may not have the appropriate AI literacy skills required for the future of work; additionally, universities are perceived to be slow in preparing graduates with the necessary skills for future work in an AI-era [DEC (2024)]. According to the survey, 58% of students feel they lack sufficient AI knowledge and skills, reflecting similar

concerns among business executives. The survey also reveals how students are starting to inappropriately rely on GenAI for routine tasks like information retrieval and analysis, paralleling how they might offload data processing and decision making to AI once they enter the workforce. These trends highlight a broader need for continuous AI literacy training, equipping both students and business leaders with the knowledge and skills necessary to navigate and leverage AI-driven environments.

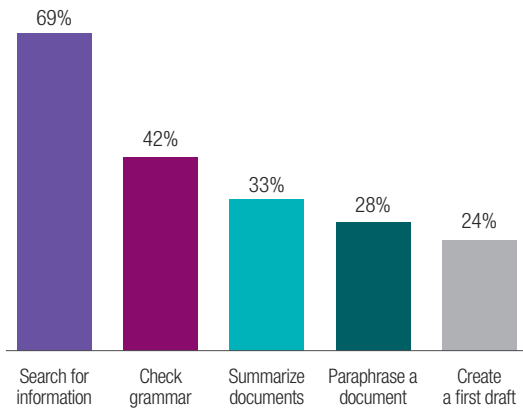
It is critical to emphasize that human oversight must guide AI in contexts requiring nuanced understanding, such as creative brainstorming, complex and contextual problem solving, and ethical decision making. In such scenarios, AI functions best as a supplement to human judgment, not a replacement [Chen

et al. (2023), Ng et al. (2024)]. Studies highlight that AI can mitigate human biases and reduce cognitive load by offering data-driven insights, but must be paired with human oversight to ensure context-aware, creative, and ethical decision making [Ng et al. (2024), Dahri et al. (2024), Chen et al. (2023), Essel et al. (2024), Hao et al. (2024)]. In this article, we will explain how to achieve a desirable outcome by avoiding the dangers posed by inappropriate cognitive offloading by appropriately leveraging metacognition.

2. THE DANGER: INAPPROPRIATE COGNITIVE OFFLOADING

Cognitive offloading refers to the process by which individuals delegate tasks that require memory, computation, or decision making to external tools, thereby reducing their cognitive load. The DEC Global AI Student Survey reveals that students are increasingly using GenAI tools for routine tasks such as information retrieval, summarizing, and drafting (Figure 1). By offloading these cognitive tasks to AI, students can theoretically focus more on higher-order thinking and creativity. The survey further highlights that students are utilizing AI not just for academic tasks but also for career-related activities, including drafting resumes and cover letters, practicing for mock interviews, and receiving career recommendations (Figure 2). This demonstrates the broader role of AI in reducing cognitive load across both academic and professional contexts, allowing users to allocate cognitive resources to more complex decision making and strategic planning.

Figure 1: How students usually use AI tools



Source: DEC (2024)
 Note: Students were asked to select all that applied

Figure 2: AI use cases in higher education (ranked by student perception)



Source: DEC (2024)
 Notes: Figures indicate the percentage of students who viewed a use case for AI positively.

Ideally, such cognitive offloading should enable students to focus on higher-order thinking, delegating routine tasks to AI. Similarly, in the business world, executives could adopt AI for routine tasks like simple data processing, market analysis, and operational decision making. By offloading such tasks to AI, the humans involved could dedicate more cognitive resources to more creative problem solving and strategic thinking.

The danger here, however, is that excessive dependence on AI could reduce human cognitive engagement, leading to poor decision making and diminished problem-solving abilities [Ng et al. (2024), Chen et al. (2023)]. While cognitive offloading to AI can free up mental resources for more complex tasks, overreliance on AI tools could lead to a decline in critical thinking and problem-solving skills. Students who frequently use AI for tasks like summarizing or drafting may bypass the deeper cognitive engagement necessary to fully understand the material, ultimately hindering their learning. Dell'Acqua et al. (2023) find that AI can improve human productivity and quality when performing complex tasks, enabling workers to perform at a higher level by offloading routine and repetitive activities to AI. However, it is difficult for humans to understand what tasks AI can and cannot do effectively [see the “jagged technological frontier” in Dell'Acqua et al. (2023)]. For instance, using AI for career-related tasks like resume creation may diminish opportunities for self-reflection and personal growth, potentially stunting the development of key professional skills and insights. Similarly, business leaders may begin to overlook important contextual factors if they depend solely on AI outputs without engaging in critical reflection on the nuances of their industry. Such an overreliance on AI can impair long-term cognitive skills and result in suboptimal decisions.

The key will be to think of AI as an assistant (or “co-pilot”) rather than a decision maker, with humans retaining control of final judgments [Ng et al. (2024)]. Randazzo et al.'s (2024) study conceptualizes three types of human-AI knowledge co-creation: “fused co-creation” (cyborgs), where professionals fully integrate AI into their workflows; “directed co-creation” (centaurs), where tasks are divided between humans and AI based on their strengths; and “abdicated co-creation” (self-automators), where professionals rely entirely on AI without developing new skills. When and where each form of co-creation is appropriate should drive how AI is used; in other words, form should follow function. Yet, students and executives alike are increasingly mis-using AI. Part of the problem is that they lack a basic understanding of AI (i.e., they lack AI literacy). An even bigger part of the problem, however, may be that humans are trying to apply AI without thinking

carefully about what they are doing and how they are doing it, which are prerequisites for properly introducing AI into decision making.

3. THE SOLUTION: METACOGNITION

This possibility highlights the need for metacognitive awareness, or the ability to reflect on one's thinking and decision making strategies. Metacognition refers to the ability to execute a sequence of strategies, employ heuristics that lead to success on a task, and explicitly self-regulate one's behavior during complex tasks [Flavell (1979), Hennessey (1999)]. It involves conscious awareness and control over one's cognitive processes, enabling individuals to plan, monitor, and adjust their approach to problem solving. For instance, an executive might be aware of their tendency to send emailed responses to complex situations without adequately thinking about it first, and make sure to sleep on it before responding. This concept excludes basic learning strategies like making inferences or summarizing text, foundational problem analysis such as defining entities and testing solutions, and general self-regulative behaviors like seeking clarification or offering alternative explanations. Instead, metacognitive awareness focuses on higher-order thinking skills that allow individuals to navigate complex tasks effectively and optimize their performance.

Metacognitive awareness enables individuals to assess when and how to use AI effectively, preventing overreliance on technology while ensuring AI outputs are properly evaluated. According to DEC (2024), 55% of students express concerns about becoming too dependent on AI. Specifically, students worry that overreliance on AI in teaching and learning could attenuate their learning experiences and 52% worry that it would negatively impact their academic performance. The survey's authors state that, “students do not want to become overreliant on AI, and they do not want their professors to do so either.” While students recognize the benefits of incorporating AI into education, they also perceive the risks of overdependence. Additionally, concerns arise that excessive use of AI in teaching could lead students to question the quality of their education and the fairness of AI-driven evaluations, especially if educators are not actively involved in the process. Although there is limited research on how these perceptions might translate into the workplace, executives should be mindful that similar concerns about overreliance on AI may exist in professional settings as well. AI's rapid processing capabilities could yield creative solutions in the workplace, but human leaders must actively regulate its use by exercising metacognition [Chen et al. (2023)].

This highlights the need for metacognitive strategies that encourage users to remain engaged in decision making, balancing AI assistance with human judgment. Metacognitive control is crucial in this context, allowing students and business leaders to reflect on the limitations of AI, particularly in interpreting context-specific variables. For example, executives using AI to draft reports must consciously evaluate how AI-generated outputs align with their specific needs [Ng et al. (2024)], actively reflecting on the appropriateness of AI's role in each decision. Metacognitive skills are essential for such management.

The DEC survey also reveals student concerns about the ethical use of AI and the potential biases embedded in AI systems, which further emphasizes the need for a critical thinking approach to AI-enhanced decision making (Figure 3). AI, while capable of processing vast amounts of data, can still produce biased or incomplete insights. Metacognitive awareness enables executives to question AI outputs, examine their assumptions, and ensure that decisions are made ethically and contextually. This balance between human oversight and AI integration ensures that cognitive offloading does not lead to passive decision making but instead enhances overall cognitive performance [Chen et al. (2023), Essel et al. (2024)].

Developing metacognitive skills allows leaders to assess AI-generated insights not just for their accuracy but also for their alignment with organizational values and goals, continuously evaluating its impact on strategy, ethics, and business performance. Without such a reflective approach, AI risks becoming a tool for automation rather than augmentation. Consequently, fostering metacognitive awareness in AI-enhanced environments is essential for ensuring that decisions remain human-centered, ethically sound, and strategically effective.

4. METACOGNITIVE STRATEGIES FOR EFFECTIVE AI INTEGRATION

To harness the full potential of AI while avoiding inappropriate cognitive offloading, individuals and organizations should develop and apply metacognitive strategies. In the context of AI, metacognitive strategies enable users to critically assess when and how to utilize AI tools effectively. This involves a series of steps that include:

4.1 Environmental awareness

- **Contextual understanding:** recognizing the environment, tools, and constraints that may impact task completion, including resource availability and organizational readiness. Being aware of the resources available and any limitations helps in planning effectively.
- **Sensory information processing:** actively gather and interpret information from your surroundings to inform your decisions and encourage open communication within teams to share insights.

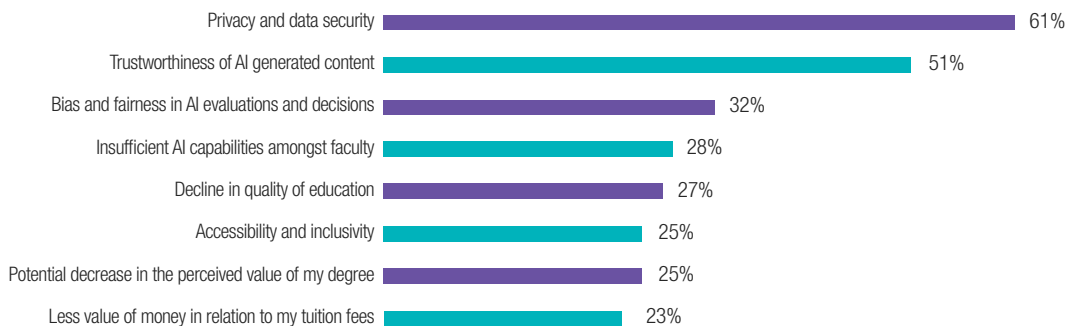
4.2 Planning and goal setting

- **Define objectives:** clearly outline what you aim to achieve before engaging with AI, considering the context and available resources. Set measurable goals and identify key performance indicators (KPIs) to track progress.
- **Determine task appropriateness:** assess which tasks are suitable for AI assistance and which require human insight.

4.3 Active monitoring

- **Self-questioning:** continuously ask yourself if AI outputs make sense. For example, "Is this recommendation logical given the data?"

Figure 3: Student concerns about their universities' use of AI



Source: DEC (2024)

Note: Figures indicate the percentage of students who expressed concern regarding AI usage within a specified topic.

- **Awareness of biases:** be vigilant about potential AI biases and your own cognitive biases that might affect interpretation.
- **Guard against latent persuasion:** be aware that AI tools can subtly shape the opinions you express and ultimately believe – a phenomenon known as latent persuasion [Sparks et al. (2024)]. Actively question whether AI is unintentionally steering your beliefs and ensure your conclusions are grounded in independent analysis and critical thinking.
- **Team feedback mechanisms:** implement regular team discussions to review AI outputs collectively, fostering a collaborative approach to monitoring and evaluation.

4.4 Critical evaluation

- **Continuous evaluation:** while interacting with AI, continuously assess whether the outputs make sense and align with your objectives and the level of quality you expect.
- **Cross-verification:** validate AI outputs with additional sources or data when possible.
- **Outcome reflection:** after decisions are made, reflect on the role AI played and whether it enhanced the decision making process.

4.5 Adaptive learning

- **Feedback integration:** use past experiences to inform future interactions with AI. This should be an iterative exercise with constant reflection. Document lessons learned and share them across the organization to promote collective learning.
- **Continuous education:** stay updated on AI developments to understand new capabilities and limitations, while reflecting its impact on decision making and problem solving.

5. ADDRESSING THE AI LITERACY GAP

Exercising metacognitive awareness is necessary but not sufficient by itself. Students and executives alike need to better understand what AI is, what it can do, and what it cannot. Gaps in AI literacy remain widespread. For instance, DEC (2024) reveals a significant gap in AI literacy, with 58% of students feeling underprepared for the future of work. These findings closely parallel the challenges faced in the corporate world, where many executives also acknowledge the need to bridge gaps in AI literacy within their organizations. Executives should focus on developing AI literacy along with metacognition, not only using AI but also reflecting on its limitations and capabilities [Ng et al. (2024)].

Box 1: Applying metacognitive strategies in practice

For instance, when using AI to generate a market analysis report, you should:

- **Plan:** define what insights you need and decide which sections AI can assist with.
- **Monitor:** as the AI generates content, regularly check for accuracy and relevance.
- **Evaluate:** critically assess the final output for any inconsistencies or gaps.
- **Adapt:** note any issues encountered and adjust your approach for next time.

There is a clear expectation from students that universities should play a central role in developing the skills necessary to manage AI effectively, something that business executives also need. Cultivating metacognitive strategies to maximize AI's potential in decision making will be crucial for success both before and after entering the workplace [Chen et al. (2023)]. Just as universities are being urged to enhance AI education, companies should also invest in ongoing AI literacy programs for their workforce. By offering professional development opportunities focused on AI, businesses can ensure that their employees remain competitive and proficient in using AI tools, thereby enabling its workforce to adapt to advances in AI.

6. CONCLUSION

As AI continues to reshape decision making across industries, the development of models like ChatGPTo1 highlights the rapid advancements in AI capabilities [OpenAI (2024)]. This new model, designed to reason through complex tasks and solve multi-step problems in areas as diverse as math, coding, and science, exemplifies how AI might become more sophisticated in mimicking human-like thought processes. The new model can evaluate multiple options before responding, making it significantly better at handling complex problems compared to previous models. While this progress shows AI's immense potential to transform problem solving and innovation, it also underscores the critical need for metacognitive awareness and human oversight in AI-driven environments given the ongoing need to adapt to technological advances.

This example emphasizes why human-in-the-loop decision making is more important than ever. Despite ChatGPTo1's ability to perform better than previous models, particularly in technical tasks, its outputs must still be carefully evaluated

by humans, especially in ambiguous or high-risk scenarios. AI tools like ChatGPT are more capable but still cannot fully understand the ethical or contextual complexities that may arise in business decisions. While AI can enhance problem solving and improve efficiency, it should nevertheless serve as a thought partner rather than a replacement for human judgment.

The risk of inappropriate cognitive offloading may also be exacerbated by these rapid advancements. As AI becomes more proficient, there is a growing temptation to offload more critical tasks to these systems. However, overreliance on AI could lead to diminished cognitive engagement and poorer decision making over time. Just like current students, business leaders must remain actively involved, applying metacognitive strategies to reflect on AI outputs, question assumptions, and ensure that decisions are made with a thorough understanding of the broader context.

The same advances also support the need for AI literacy among both students and executives. As AI systems become more complex, the ability to critically evaluate their outputs and understand their limitations becomes even more essential. Organizations must prioritize ongoing AI training and education, ensuring that their workforce is not only proficient in using AI but also equipped to oversee and guide AI in a way that aligns with ethical standards and business goals.

Throughout these processes, students and employees alike should remain involved in decisions regarding AI integration. DEC (2024) found that students wanted to help shape AI's role in education and the workplace. This mirrors how businesses should involve leadership teams in developing AI training strategies to avoid the risks of AI overreliance [Ng et al. (2024), Essel et al. (2024)]. By engaging executives and managers in AI training and decision making, organizations can ensure a more holistic approach to AI integration. Beyond enhancing the effectiveness of AI initiatives, this approach also builds a sense of ownership and accountability in how AI is used to shape business strategies and operations.

The rapid development of AI clearly demonstrates its growing capability to assist in decision making and problem solving. However, with these advancements comes the pressing need for careful human oversight. Business leaders must ensure that AI tools are used thoughtfully, balancing the convenience of offloading tasks with the necessity of staying engaged in critical decision making. By developing metacognitive skills and maintaining an active role in overseeing AI outputs, leaders can ensure that AI serves to enhance, rather than replace, human judgment, creativity, and ethical responsibility.

REFERENCES

- Chen, B., X. Zhu, and H. F. Díaz del Castillo, 2023, "Integrating generative AI in knowledge building," *Computers and Education: Artificial Intelligence*, 5
- Dahri, N. A., N. Yahaya, W. M. Al-Rahmi, A. Aldraiveesh, U. Alturki, S. Almutairy, ... and R. B. Soomro, 2024, "Extended TAM based acceptance of AI-Powered ChatGPT for supporting metacognitive self-regulated learning in education: A mixed-methods study," *Heliyon*, e29317
- DEC, 2024, "The Digital Education Council Global AI student survey (2024)," <https://tinyurl.com/4fk7xuhf>
- Dell'Acqua, F., E. McFowland III, E. R. Mollick, H. Lifshitz-Assaf, K. Kellogg, S. Rajendran, ... and K. R. Lakhani, 2023, "Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality," *Harvard Business School Technology & Operations Management unit working paper no. 24-013*
- Di Battista, A., S. Grayling, E. Hasselaar, T. Leopold, R. Li, M. Rayner, and S. Zahidi, 2023, "Future of jobs report 2023," *World Economic Forum*, <https://tinyurl.com/25z6pf86>
- Essel, H. B., D. Vlachopoulos, A. B. Essuman, and J. O. Amankwa, 2024, "ChatGPT effects on cognitive skills of undergraduate students: receiving instant responses from AI-based conversational large language models (LLMs)," *Computers and Education: Artificial Intelligence*, 6, 100198
- Flavell, J. H., 1979, "Metacognition and cognitive monitoring: a new area of cognitive-developmental inquiry," *American psychologist* 34:10, 906
- Hatzius, J., J. Briggs, D. Kodnani, and G. Pierdomenico, 2023, "The potentially large effects of artificial intelligence on economic growth (Briggs/Kodnani)," *Goldman Sachs Economic Research*, March 26
- Hennessey, M. G., 1999, "Probing the dimensions of metacognition: implications for conceptual change Teaching-Learning," Paper presented at the Annual Meeting of the National Association for Research in Science Teaching, Boston
- Jones, S., and O. Olson, 2024, "Preparing MBA students for leadership in GenAI environment: what educators need to know," *Effective Executive* 27:2, 49-55
- Kabashkin, I., B. Misnevs, and O. Zervina, 2023, "Artificial intelligence in aviation: new professionals for new technologies," *Applied Sciences* 13:21, 11660
- Lakhani, K., and A. Ignatius, 2023, "AI won't replace humans - but humans with ai will replace humans without ai," *Harvard Business Review*, <https://tinyurl.com/38fbtrct>
- Ng, D. T. K., C. W. Tan, and J. K. L. Leung, 2024, "Empowering student self-regulated learning and science education through ChatGPT: A pioneering pilot study," *British Journal of Educational Technology*
- OpenAI, 2024, "Introducing OpenAI o1-preview," <https://tinyurl.com/5unaf5a2>
- Randazzo, S., and H. Lifshitz-Assaf, K. Kellogg, F. Dell'Acqua, E. R. Mollick, K. R. Lakhani, 2024, "Cyborgs, centaurs and self automators: human-genai fused, directed and abdicated knowledge co-creation processes and their implications for skilling," SSRN, <https://tinyurl.com/yc4zxr6>
- Sparks, J. R., T. M. Ober, C. Tenison, B. Arslan, I. Roll, P. Deane, D. Z. Rivera, R. M. Gooch, and T. O'Reilly, 2024, "Opportunities and challenges for assessing digital and AI literacies," *ETS research Institute*, <https://tinyurl.com/5fwah72v>
- Suleman, R. M., R. Mizoguchi, and M. Ikeda, 2016, "A new perspective of negotiation-based dialog to enhance metacognitive skills in the context of open learner models," *International Journal of Artificial Intelligence in Education* 26, 1069-1115



REGULATION

104 Mapping GenAI regulation in finance and bridging the gaps

Nydia Remolina, Assistant Professor of Law, and Fintech Track Lead, SMU Centre for AI and Data Governance, Singapore Management University

112 Board decision making in the age of AI: Ownership and trust

Katja Langenbacher, Professor of Civil Law, Commercial Law, and Banking Law, Goethe University Frankfurt

122 The transformative power of AI in the legal sector: Balancing innovation, strategy, and human skills

Eugenia Navarro, Lecturer and Director of the Legal Operations and Legal Tech Course, ESADE

129 Remuneration on the management board in financial institutions: Current developments in the framework of supervisory law, labor law, behavioral economics and practice

Julia Redenius-Hövermann, Professor of Civil Law and Corporate Law and Director of the Corporate Governance Institute (CGI) and the Frankfurt Competence Centre for German and Global Regulation (FCCR), Frankfurt School of Finance and Management

Lars Hinrichs, Partner at Deloitte Legal Rechtsanwaltsgesellschaft mbH (Deloitte Legal) and Lecturer, Frankfurt School of Finance and Management

MAPPING GenAI REGULATION IN FINANCE AND BRIDGING THE GAPS

NYDIA REMOLINA | Assistant Professor of Law, and Fintech Track Lead, SMU Centre for AI and Data Governance,
Singapore Management University

ABSTRACT

Generative artificial intelligence (GenAI) is rapidly reshaping the financial services sector by introducing new avenues for innovation, efficiency, and profitability. GenAI systems, including models like “generative adversarial networks” (GANs) and “transformers”, can autonomously generate content such as synthetic data, trading strategies, and fraud detection insights, transforming traditional financial operations. However, these advancements come with new challenges, particularly in ensuring that GenAI is deployed ethically, securely, and in compliance with evolving regulatory frameworks. Current financial regulations, such as those governing anti-money laundering (AML), market integrity, financial consumer protection, among others, were originally designed for human-driven processes and do not fully address the complexities introduced by AI systems. While some jurisdictions, such as the E.U., Singapore, the U.S., and China, have launched AI regulatory initiatives, frameworks specifically tailored to the financial services industry are still a work in progress. This article seeks to provide an overview of these differing regulatory landscapes while raising awareness of the gaps that financial institutions and regulators should address to bridge in the responsible adoption of GenAI in the financial services sector.

1. INTRODUCTION

Generative artificial intelligence (GenAI) is rapidly transforming the financial services sector, ushering in new opportunities for innovation, efficiency, and profitability [Teresa (2023)]. GenAI refers to a class of artificial intelligence systems that can create new, original content or data by learning from existing data patterns. Using advanced models like “generative adversarial networks” (GANs) and “transformers”, generative AI can produce text, images, audio, and other types of content that mimic human-like creativity and decision making [Foster (2022)]. In finance, GenAI is used for applications such as synthetic data generation, algorithmic trading strategies, fraud detection, and personalized financial services [Lee et al. (2024), Ramdurai and Adhithya (2023)]. Its ability to autonomously generate content or simulate scenarios sets it apart from traditional AI models that simply analyze or classify data. Hence, these technologies promise to reshape how financial institutions operate, making processes faster and more accurate while reducing costs

[Wang (2023)]. However, with these advancements come new challenges, particularly in ensuring that GenAI systems are deployed ethically, securely, and within the bounds of regulatory frameworks that jurisdictions around the world have developed in the last few years to mitigate the risks of AI systems (predictive and generative) [for more on the differences between predictive AI and GenAI, see Hermann and Puntoni (2024), Harrington (2024)].

Despite the rapid uptake of artificial intelligence (AI) in finance, regulatory frameworks have struggled to keep pace [Roberts et al. (2024)]. Many existing regulations, such as those governing anti-money laundering (AML), data privacy, market integrity, financial stability, consumer protection, were designed for human-centered processes and may not fully address the complexities introduced by automated systems [Remolina (2024)]. Additionally, regulations specific to AI offer broad guidelines but often lack the granularity and a sector-specific approach needed for the unique applications of AI

in the financial services sector. Additionally, GenAI has just recently entered into the policy and regulatory conversation for financial regulators in some jurisdictions.

Indeed, Singapore, the E.U., the U.S., and China have each launched initiatives to regulate AI, and some of them to regulate GenAI. However, we are still at an early stage in these developments and none offer a framework tailored to the financial services industry with appropriate enforcement mechanisms to tackle the new risks created by GenAI. In finance, where trust, transparency, and accountability are paramount, these gaps pose real risks that threatens financial consumer protection and even the stability of the financial sector.

This article maps the characteristics of the current regulatory models for GenAI in finance, from some first-mover jurisdictions such as the U.S., the E.U., Singapore, and China, identifying where regulations succeed, where they fall short, and what gaps need to be addressed to ensure safe and ethical AI adoption. By analyzing various jurisdictions and their regulatory approaches, this article seeks to provide an overview of the regulatory landscape while raising awareness of the gaps that financial institutions and regulators should address to bridge the gaps in the responsible adoption of GenAI in the financial services sector.

2. THE STATE OF DEVELOPMENT AND IMPLEMENTATION OF GenAI IN FINANCE

GenAI is gaining significant traction in the financial services sector. Although GenAI's implementation is still at an experimental stage, it has the potential of transforming the way financial institutions operate and interact with both consumers and markets [Aldasoro et al. (2024)]. These AI systems, which can create data, content, and predictions autonomously, are being integrated into various areas such as algorithmic trading, fraud detection, customer service, and personalized financial planning. As the demand for real-time decision making and advanced predictive capabilities grows, GenAI is positioned to play a pivotal role in the future of finance.

For instance, financial institutions such as JP Morgan Chase is using GenAI to enhance fraud detection by creating synthetic transaction data. This synthetic data is fed into machine learning models to train the system without compromising real customer information, which enables better fraud detection and risk management [Trinh (2024)]. Likewise, Mastercard utilizes GenAI to combat fraud by developing AI-generated models that can simulate fraudulent activities and predict

patterns of suspicious behavior. This initiative uses AI to create fraud prevention models. These models analyze transactional data in real-time, allowing Mastercard to reduce false positives in fraud detection, improve customer experience, and lower operational costs to the point that Mastercard has reportedly decreased false positives during the detection of fraudulent transactions against potentially compromised cards by up to 200%, and increased the speed of identifying merchants at risk from – or compromised by – fraudsters by 300% [Mastercard (2024)].

Ant Financial, one of the world's largest digital payment platforms, uses GenAI for both risk assessment and customer service [Fan (2024), Asian Banker (2024)]. The company employs AI to create detailed risk profiles for users, leveraging data from various sources to make quick and accurate credit decisions. Maxiaocai, an AI agent, offers users expert-level financial services, customized market insights, simplified complex financial concepts, and tailored investment advice. The AI personal financial manager can generate visual summaries of financial reports, highlighting essential information, and translate intricate financial terminology into easily comprehensible language [Refna (2024)]. Since its public testing began in early 2024, Maxiaocai is claimed to have garnered 70 million monthly active users as of August 2024, with 45% residing in cities below the third tier. The platform now connects with more than 200 financial institutions, including asset management companies and securities firms, as well as over 15,000 financial content creators [Refna (2024)].

Also, Zest AI, a fintech company focused on credit underwriting, uses GenAI in lending decisions. The AI model analyses and generates alternative data, helping lenders assess creditworthiness more accurately without relying solely on traditional credit scores [Deepchecks Community (2024)]. Zest AI's generative models have increased loan approval rates for historically underserved groups by 15-20% [Becky (2024)].

3. THE RISKS OF GenAI IN THE CONTEXT OF FINANCIAL SERVICES

GenAI presents existing concerns related to AI, such as lack of transparency and explainability, fairness challenges, data protection issues, while also introducing new challenges that demand attention from policymakers and the financial services sector. A prominent issue currently discussed in the industry and academia is hallucinations. In the context of financial

services, this would be a “financial hallucination” [Remolina (2024)], where GenAI produces information that is incorrect or misleading [Weidinger (2022), Wachter (2024)]. Hallucinations can lead to inappropriate risk assessments or incorrect advice through AI-supported chatbots, undermining public trust in both the AI systems and the financial institutions using them.

Data privacy and protection are also significant concerns with GenAI, especially in finance, a highly regulated industry. These models are typically trained on large datasets, which may include sensitive financial information. The use of publicly accessible AI platforms within financial institutions can increase the risk of inadvertently exposing confidential data. Many AI platforms do not guarantee data protection, leaving financial institutions vulnerable to breaches. This issue is especially pressing for smaller institutions that lack the resources to develop in-house AI models, which would offer better control over data security [Remolina (2024)].

Fairness is another critical issue with GenAI, particularly when it is used in financial decision making processes like credit scoring. If the training data is biased, the AI's outputs will reflect and potentially amplify those biases, leading to discriminatory outcomes. This is especially problematic in lending markets, where biased AI systems could restrict access to credit for certain groups. Although some regulators, such as the Monetary Authority of Singapore (MAS), encourage financial institutions to assess algorithmic credit scoring through the Veritas Initiative, these recommendations are not mandatory¹ and do not fully address the specific challenges posed by GenAI given that they were proposed in the context of predictive AI [Remolina (2022)].

GenAI also impacts systemic risk in the financial services sector. The widespread and interconnected use of AI increases the risk of market instability, particularly due to the procyclicality of AI-driven decisions and the speed at which they are made. Overreliance on AI-generated reports could result in herd behavior, leading to mispricing and market imbalances [Shabsigh and Boukherouaa (2023)]. Moreover, the concentration of foundational AI model providers could create new concentration problems in a complex new financial infrastructure, as many of these providers operate beyond the reach of financial regulators [Remolina (2023)].

GenAI also raises intellectual property concerns, particularly regarding copyright infringement [Lemley (2024)]. Many GenAI models are trained on proprietary financial analyses and reports without proper authorization, potentially violating copyright laws. Some jurisdictions are exploring licensing solutions and copyright guidelines to address these legal challenges [Samuelson (2023)].

Lastly, the problem of value alignment is significant in GenAI. In finance, ensuring that AI-generated decisions align with human values and ethical standards is crucial. If AI systems generate overly risky or deceptive financial strategies, the consequences could be disastrous, undermining trust in financial institutions and threatening the stability of the financial system.

4. PROBLEMATIC CHARACTERISTICS OF THE EARLY REGULATORY MODELS THAT IMPACT GenAI IN FINANCE

Regulatory frameworks specifically addressing GenAI in finance remain underdeveloped, and the approaches taken by jurisdictions like the U.S., Singapore, the E.U., and China vary significantly, while sharing some similarities. This section compares the main characteristics of the current regulatory models in these regions and explores their impact on the financial services sector.

In the U.S., regulatory oversight for AI in finance is fragmented and sector-specific. There is no centralized AI law governing its use in financial services. Instead, the U.S. relies on existing regulations such as the Equal Credit Opportunity Act (ECOA), which requires fairness in credit decisions, including those made using automated systems [Gillis (2022)]. Additionally, data privacy laws like the California Consumer Privacy Act (CCPA) aim to protect consumer data in AI-driven processes. Financial regulators such as the Federal Trade Commission (FTC) and Securities and Exchange Commission (SEC) also play a role in monitoring the use of AI in financial services, particularly in ensuring transparency, mitigating fraud, and protecting consumers. However, these frameworks do not directly address the unique risks posed by GenAI, such as model explainability or the mitigation of biases that may arise from AI-generated content.

¹ Nonetheless, there is an expectation in Singapore that the industry should comply with these recommendations.

In contrast, several policy initiatives have positioned Singapore as one of the leading advocates of AI governance worldwide. In Singapore, the approach to AI governance and regulation is based on non-mandatory tools for the private sector to develop ethical and responsible AI systems, and a cooperative effort between regulators and the private sector. By 2019, Singapore had launched initiatives such as the Model AI Governance Framework; an international and industry-led advisory council on the ethical use of AI and data;² and a research program on the governance of AI, ethics, and data-use established through the Centre for AI and Data Governance at the Singapore Management University [Goh and Remolina (2020)].

The Model AI Governance Framework was published as a guide for organizations to address key ethical and governance issues when deploying AI technologies [PDPC (2020)]. A second edition of the model framework was launched by the Minister for Communications and Information at the World Economic Forum Annual Meeting in 2020. Both editions identify two sets of ethical principles for the responsible adoption of AI in the private sector, namely: decisions made by AI should be explainable, transparent, and fair; and AI systems should be human-centric. The model framework is complemented with the Implementation and Self-Assessment Guide for Organizations (ISAGO), which aims to help organizations decide how their AI governance practices can align with the “model framework”. ISAGO provides a set of questions and practical examples to enable organizations to assess their AI governance practices against the model framework [WEF and IMDA (2020)].

In 2023, the AI Verify Foundation was launched to develop AI testing tools for the responsible use of AI [Gurreea-Martinez and Remolina (2024)]. In relation to sector-specific strategies, MAS published, in 2018, a guide on principles to promote fairness, ethics, accountability, and transparency (FEAT) in the use of AI in the financial sector [MAS (2018)]. In addition, MAS launched the Veritas initiative to translate into practice the FEAT principles in specific AI use cases in the financial services sector; for instance, by assessing discrimination and fairness issues in algorithmic credit scoring [MAS (2021)]. Furthermore, in 2023, the Info-communications Media Development Authority (IMDA) unveiled the GenAI evaluation sandbox to test AI governance in concrete GenAI use cases

“

While significant strides have been made in regulating GenAI in finance, there remain substantial gaps in the current frameworks.

”

and the AI Verify Foundation and IMDA published in 2024 a proposed “AI Model Governance Framework for Generative AI” to mitigate the risks enforced and newly created by this type of AI [IMDA and AI Verify Foundation (2024)]. This proposal advocates for a practical, risk-based approach to evaluating GenAI, focusing on six core areas: accountability in AI development, data usage in model training, model development and deployment, third-party evaluations, research on safety and alignment, and using AI to promote public good. The paper also called for more global cooperation to establish a unified platform for GenAI governance.

The E.U. AI Act categorizes AI systems based on risk, with high-risk applications in finance, such as credit scoring, facing stringent oversight. This includes requirements for explainability, transparency, and risk management [European Parliament (2023)]. Additionally, the General Data Protection Regulation (GDPR) imposes data protection requirements on AI systems, ensuring that personal data is handled ethically and responsibly.

4.1 A fragmented approach

There is no comprehensive, globally accepted regulatory framework specifically for AI in finance. Instead, jurisdictions apply existing regulations from areas such as data privacy or data protection (e.g., GDPR), financial integrity (e.g., anti-money laundering regulations), cybersecurity, consumer protection laws, and the unintegrated approaches to GenAI governance that do not necessarily consider its coexistence with all the ecosystem of multiple regulations.

² Singapore’s Advisory Council on the Ethical Use of AI and Data was established on August 30th, 2018. The 11 Advisory Council members are from diverse backgrounds and comprise of international leaders in AI, including from big technology companies, advocates of social and consumer interests, and local companies. The Advisory Council assists the authorities in engaging with stakeholders to support the development of AI governance through issuing advisory guidelines, practical guides, and codes of practice for voluntary industry adoption. IMDA, 2019, “ANNEX A: Council Members of the Advisory Council on the ethical use of AI and data.”

This patchwork approach has resulted in a fragmented regulatory landscape, with each jurisdiction, and within a jurisdiction, different regulators developing their own rules that dictate or recommend (in the cases of non-mandatory approaches) how GenAI can be implemented in financial services, leading to a lack of uniformity. For instance, the U.S. Equal Credit Opportunity Act (ECOA) indirectly requires fairness in automated decision making, ensuring that AI models do not produce discriminatory outcomes, but similar guidelines are not uniformly adopted across all regions. Data protection laws like the GDPR in the E.U. and the California Consumer Privacy Act (CCPA) in the U.S. further complicate this fragmented approach, as financial institutions must navigate different compliance requirements when deploying GenAI across borders. This fragmented approach creates regulatory uncertainty and increases compliance costs for financial institutions operating globally. Without an effort to coherently integrate GenAI regulation into this complex ecosystem of rules, financial institutions face challenges in aligning their AI systems with the varying expectations of regulators, particularly in areas such as bias mitigation, explainability, and data protection.

4.2 Homogenization of GenAI regulation for all sectors

One of the major gaps in the current regulatory landscape is the lack of a sector-specific approach to GenAI in finance. While general AI regulations such as the E.U.'s AI Act provide broad guidance, they do not address the unique complexities of GenAI in financial services. Financial markets are highly sensitive to issues such as data security, risk management, and market manipulation, which require a specialized regulatory framework.

Additionally, even within the financial services sector, GenAI – and AI – use cases do not create the same risks. For example, AI-generated trading strategies or automated lending decisions may have direct and immediate impacts on market stability, consumer welfare, and systemic risk. However, GenAI for fraud detection does not pose an immediate risk to financial stability while algorithmic trading could pose a greater risk in this area. The existing regulatory models do not fully account for these sector and subsector-specific risks, leaving financial institutions exposed to potential legal and reputational consequences. A sector-specific approach would provide more targeted guidelines and enforcement mechanisms to ensure that GenAI is deployed safely and ethically in financial contexts.

Singapore's approach is one of the first in trying to provide a more tailored approach to the financial services sector in the governance considerations for the use and deployment of GenAI through project MindForge. Project MindForge is driven by the Veritas Initiative and examines the risks and opportunities of GenAI technology for financial services [MAS (2023)]. It aims to develop a clear and concise framework on the responsible use of GenAI in the financial services industry. In phase one, the consortium has developed a comprehensive GenAI risk framework, with seven risk dimensions identified in the areas of:

1. Accountability and governance
2. Monitoring and stability
3. Transparency and explainability
4. Fairness and bias
5. Legal and regulatory
6. Ethics and impact
7. Cyber and data security

4.3 Self-regulation as the main risk-mitigation tool

The reliance on self-regulation is another significant aspect of the regulatory models for GenAI in finance. The E.U. has faced criticism for allowing financial institutions and AI developers to self-regulate in certain areas, leading to concerns about weak oversight and the potential for harm that is not immediately tangible, such as bias or financial losses due to faulty AI systems. Self-certification processes, while intended to encourage innovation, may not provide sufficient safeguards against issues like financial hallucinations or systemic risks.

Similarly, China has adopted a self-regulatory model with its “Interim Measures for the Management of Generative AI Services”, which requires AI systems to undergo security assessments and adhere to content governance rules. However, these measures focus more on public safety and political considerations rather than the specific risks associated with GenAI in financial services. While China's approach emphasizes transparency and ethical AI use, it lacks the financial sector-specific focus necessary to address the full range of risks posed by GenAI in finance.

4.4 Materiality and risk assessment

A key challenge in regulating GenAI in finance is the need for clear guidelines on materiality and risk assessment. Financial institutions must assess the material impact of GenAI

models on decision making processes, particularly in areas like lending, trading, and fraud detection. However, current regulatory frameworks often lack concrete standards for how to measure the risks associated with AI-generated outputs, making it difficult for institutions to conduct comprehensive risk assessments.

For example, the potential for AI models to generate misleading financial reports or biased lending decisions requires financial institutions to develop new tools and methodologies for assessing the material risks posed by these systems. Regulatory bodies need to provide clearer guidelines on how to quantify and mitigate the risks associated with GenAI, particularly in the context of systemic risk and financial stability.

Initiatives such as Veritas or the sandboxes could serve this purpose. However, these available tools are not mandatory for the financial services sector. Additionally, when Veritas was launched and proposed, it did not consider the risks exacerbated and created by GenAI and the particularities of GenAI versus predictive AI. Additionally, the GenAI sandbox, launched in 2024 in Singapore, targets SMEs to harness the benefits of GenAI and support their innovation and digitalization journey. It is led by the Infocomm Media Development Authority (IMDA). Thus, this sandbox is not a financial regulation tool and, as such, is not specialized the financial services sector.

Other approaches, such as the E.U. AI Act, have been criticized for the overreliance on self-certification, weak oversight and investigatory mechanisms, and far-reaching exceptions for both the public and private sectors [Wachter (2024)]. The proposed liability frameworks for AI systems in the E.U. have been similarly criticized because they focus on material harm while ignoring harm that is immaterial, monetary, and societal, such as bias, hallucinations, and financial losses due to faulty AI products [Wachter (2024)].

4.5 The use of GenAI by fraudsters

Fraud is an area where GenAI poses exacerbated risks for the financial services sector. Fraudsters are increasingly using sophisticated AI to impersonate clients or legitimate representatives of financial institutions, tricking consumers or financial institutions into authorizing fraudulent payments. AI-generated scams are becoming more credible and difficult to detect, even for highly cautious consumers and financial professionals [Resistant AI (2023)]. A recent example is one where an employee of a Hong Kong-based financial services firm was deceived into transferring \$25 million after participating in a deepfake video conference call with someone posing as the company's CFO [Chen and Magramo (2024)]. Financial regulators should think about new approaches to balance the liability of financial institutions and consumers in this new era of authorized push payment fraud taking into consideration the new challenges posed by GenAI in payments systems.



5. CONCLUSION

This article maps the characteristics of the current regulatory models to GenAI in finance. It looks at a number of first-mover jurisdictions, such as the E.U., the U.S., Singapore, and China, identifying where regulations succeed, where they fall short, and what gaps need to be addressed to ensure safe and ethical AI adoption. By analyzing these regulatory approaches, this article seeks to provide a high-level overview of the regulatory models applicable to GenAI in finance and concludes that all models contribute to a fragmented approach to GenAI regulation. Moreover, apart from Singapore, the current approaches of the first movers lack sector-specific focus because they are mostly based on self-regulation tools, and do not provide clear risk assessment methodologies that measure the materiality of GenAI harms and tailor solution accordingly.

Finally, current approaches have not considered that some use cases of GenAI are developed outside regulated entities but still directly affect financial consumers and institutions, as seen with the use of GenAI for fraud. Frameworks such as fraud payment regulations may need recalibration to address the new challenges posed by this technology.

Decoding the issues present in these characteristics of the GenAI regulatory models is a first step for regulators, policy-makers and the industry to propose solutions aimed at bridge the gaps. While significant strides have been made in regulating GenAI in finance, there remain substantial gaps in the current frameworks. A more harmonized, sector-specific approach with enforcement mechanisms, and methodologies that recognize the general and undefined purpose of GenAI models is necessary to ensure that financial institutions can safely and ethically deploy these technologies.

REFERENCES

- Aldasoro, I., L. Gambacorta, A. Korinek, V. Shreeti, and M. Stein, 2024, "Intelligent financial system: how AI is transforming finance," Bank for International Settlements, paper no. 1194
- Asian Banker, 2024, "Ant Group unveils smart assistants services and foundation model technologies," <https://tinyurl.com/yz9eu2d5>
- Becky F., 2024, "MeridianLink and Zest AI expand partnership, accelerating market momentum and enabling financial institutions to drive more inclusive lending practices," MeridianLink, <https://tinyurl.com/bdez98at>
- Chen, H., and K. Magramo, 2024, "Finance worker pays out \$25 million after video call with deepfake 'chief financial officer'," CNN, February 4, <https://tinyurl.com/5x6mct5d>
- Deepchecks Community, 2024, "The revolutionary impact of generative AI in the financial sector," blog, <https://tinyurl.com/wdzv9yh5>
- European Parliament, 2023, "AI Act: a step closer to the first rules on artificial intelligence," <https://tinyurl.com/242fmyke>
- Fan F., 2024, "Tech company Ant Group launches AI life assistant," China Daily, <https://tinyurl.com/3e5nphc4>
- Foster, D., 2022, *Generative deep learning*, second edition, O'Reilly Media, Inc.
- Gillis, T., 2022, "The input fallacy," 106 *Minnesota Law Review* 1175
- Goh, Y., and N. Remolina, 2020, "The innovation of Singapore's AI ethics model framework" in *AI in 2019: a year in review*, Shanghai Institute for Science of Science, 77–78
- Gurrea-Martinez, A., and N. Remolina, 2024, "Financial regulation," in Phang, A., G. Yihan, and S. Chesterman (eds.), *Law and technology in Singapore*, Academy Publishing
- Harrington, L., 2024, "Comparison of generative artificial intelligence and predictive artificial intelligence," *AACN Advanced Critical Care* 35:2, 93-96
- Hermann, E., and S. Puntoni, 2024, "Artificial intelligence and consumer behavior: from predictive to generative AI," *Journal of Business Research* 180, 114720
- IMDA and AI Verify Foundation, 2024, "Proposed model AI governance framework for generative AI – fostering a trusted ecosystem," Info-communications Media Development Authority (IMDA) and AI Verify Foundation
- Lemley, M., 2024, "How generative AI turns copyright law upside down," *Science and Technology Law Review* 25:2
- Lee, D. K. C., C. Guan, Y. Yu, and Q. Ding, 2024, "A comprehensive review of generative AI in finance," *FinTech* 3:3, 460-478
- MAS, 2018, "Guide on principles to promote fairness, ethics, accountability and transparency (FEAT) in the use of artificial intelligence and data analytics in Singapore's financial sector," Monetary Authority of Singapore
- MAS, 2021, "Veritas initiative addresses implementation challenges in the responsible use of artificial intelligence and data analytics," Monetary Authority of Singapore
- MAS 2023, "MAS partners industry to develop generative AI risk framework for the financial sector," Monetary Authority of Singapore, <https://tinyurl.com/86b4r2f4>
- Mastercard, 2024, "Mastercard accelerates card fraud detection with generative AI technology," <https://tinyurl.com/bdcmc5x4>
- PDPC, 2020, "Singapore's approach to AI governance," Personal Data Protection Commission, <https://tinyurl.com/thhzv6x6>
- Ramdurai, B., and P. Adhithya, 2023, "The impact, advancements and applications of generative AI," *International Journal of Computer Science and Engineering* 10:6, 1-8
- Refna, T., 2024, "Ant Group launches new AI financial manager Maxiaocai," *Finance Director Europe*, <https://tinyurl.com/yh88y26p>
- Remolina, N., 2022, "The role of financial regulators in the governance of algorithmic credit scoring," SMU Centre for AI and Data Governance research paper no. 2/2022
- Remolina, N., 2023, "Interconnectedness and financial stability in the era of artificial intelligence," in Remolina, N., and A. Gurrea-Martinez (eds.), *Artificial intelligence in finance*, Edward Elgar Publishing
- Remolina, N., 2024, "Generative AI in finance: risks and potential solutions," *Law Ethics and Technology Journal* 1:1, Special Issue: The Law and Ethics of Generative AI
- Resistant AI, 2023, "FraudGPT: how AI is – and isn't – revolutionizing financial crime," April 19, <https://tinyurl.com/5v5dtbkk>
- Roberts, H., E. Hine, M. Taddeo, and L. Floridi, 2024, "Global AI governance: barriers and pathways forward," *International Affairs* 100:3, 1275-1286
- Samuelson, P., 2023, "Generative AI meets copyright," *Science* 381:6654, 158-161
- Shabsigh, G., and E-B. Boukheroua, 2023, "Generative artificial intelligence in finance," *International Monetary Fund Fintech Notes* vol. 2023, issue 006
- Teresa, W., 2023, "The benefits of generative AI for banking and financial leaders," *Actian*, October 10, <https://tinyurl.com/yj65kzke>
- Trinh N., 2024, "How JPMorgan Chase and Others are using generative AI to revolutionize banking and their financial services?" LinkedIn, <https://tinyurl.com/hfpxj2k>
- Wachter, S., 2024, "Limitations and loopholes in the EU AI Act and AI Liability Directives: what this means for the European Union, the United States, and beyond," *Yale Journal of Law and Technology* 26:3
- Wang, Y., 2023, "Generative AI in operational risk management: harnessing the future of finance," SSRN, <https://tinyurl.com/4dtybduz>
- WEF and IMDA, 2020, "Companion to the model AI governance framework – implementation and self-assessment guide for organizations," *World Economic Forum and Info-communications Media Development Authority (IMDA)*
- Weidinger, L., et al., 2022, "Taxonomy of risks posed by language models, FACCT '22: Proceedings of the 2022 ACM Conference on fairness accountability and transparency 214, 216-221

BOARD DECISION MAKING IN THE AGE OF AI: OWNERSHIP AND TRUST

KATJA LANGENBUCHER | Professor of Civil Law, Commercial Law, and Banking Law, Goethe University Frankfurt¹

ABSTRACT

This paper describes how artificial intelligence (AI) might augment board decision making and explores legal ramifications of this development. The article begins by providing a brief overview of the use of AI as a “prediction machine”² for board decisions, and then zooms in on two core characteristics that explain what corporate law requires from board decision making: that board members fully own their decisions and that board members are trusted to form business judgments, immune from judicial second-guessing. The paper makes two contributions to the debate: it rejects the notion that black-box AI may not be used for board decision making and proposes a graphic control matrix to identify low, medium, and enhanced judicial scrutiny when boards use AI to inform their decisions.

1. INTRODUCTION

Explaining human intelligence is an intriguing topic [Langenbucher (2023b)]. For some, it represents human singularity, while others emphasize the dependence of human intelligence on mechanistic operations [Glimcher (2004), Rolffs (2023), Stiehl and Marciniak-Czochra (2021)]. Whether this implies a kinship between these two forms of information processing or, conversely, whether there are fundamental differences has been discussed for hundreds of years [Hawkins (2021), Larson (2021), Nath (2009)]. Arguably, an uncontested point of departure is that machines can sometimes surpass human performance when it comes to speed and precision. From there, a pressing question follows for corporate decision making. If it is advisable for doctors, lawyers, and stock exchange traders to have certain decisions augmented by machines, does this also apply to management decisions of company directors? If so, who bears the cost if things go wrong?

The next section provides a brief overview on the use of AI as a “prediction machine” for board decisions. It reminds the reader that statistics has traditionally filled this role and hints at similarities and differences with using machine learning (ML). The following section zooms in on how corporate law frames decision making. It starts from the assumption that the law treats decision making by board members differently than decision making by officers and employees.

Against this background, the paper highlights two core characteristics. Corporate law expects board members, but not directors and employees, to fully own their decision. As a flipside of ownership, corporate law places trust in board members to form business judgments, immune from judicial second-guessing. The expectation that boards own their decisions implies that they must not abdicate their authority.

This article explores how this principle is impacted when boards enhance their decision making with an AI. It then moves on to examine how corporate law has framed ownership of a

¹ The author is also affiliated with SciencesPo in France and a regular visiting faculty at Fordham Law School in the U.S. This paper has appeared as a working paper in the ECGI/WP series. Additional thoughts (that go deeper on decision theory-related questions) will appear at Chicago Law Review Online.

² For this term see the title of Agrawal et al. (2018), as well as Agrawal et al. (2022) and Russel and Norvig (2021).

board decision when technical support tools or human experts inform board members. Rather than analogizing AI to one of these helpers, this article introduces the second dimension: trust. It claims that the standard of judicial review has moved along these two dimensions, ownership and trust. The same logic, this article suggests, applies to board decisions that integrate AI. The paper concludes with a graphic visualization of these dimensions.

2. PREDICTION MACHINES

The term “artificial intelligence” includes various implementations, ranging from logic, ML, and neural networks to large language models (LLM) and robotics [Russell and Norvig (2021)]. These correspond with a diverse set of potential use cases in corporate life. The scenario this article explores is the use of AI as a “prediction machine”. In that capacity, a board uses AI to enhance its understanding of which future events are likely to take place. Most management decisions imply predictions of that type and employing statistics to that end is a standard tool. Statisticians work on inferences about the relationship between different variables, based on a hypothesis.³ Consider the case of the management board of a bank that decides to reduce the number of brick-and-mortar branches and move towards online banking. Studies on customer preferences, possibly also their age, occupation, or place of residence, together with mobile network coverage, and the number of bank branches can inform management. An initial hypothesis might be that: the age of a customer is a core factor in driving a preference for brick-and-mortar branches.

Complementing or replacing the statistician, imagine using an AI. To train it, data on customer reactions to branch closures carried out in the past is useful. The AI furnishes patterns, such as groups of bank customers with similar preferences and reactions (clustering) [Russel and Norvig (2021)]. Its predictions about the willingness of bank customers to switch to online banking could mirror that of the statistician’s. Additionally, it might bring out unanticipated correlations. Both allow the board to react; for instance, via targeting its marketing towards specific groups.

2.1 The machine learns

One of the intriguing features of AI is its potential to learn. Instead of being provided with an input-output pair that is specified ex-ante, the AI is left to stroll through data, as it were. Its performance gets better after it has made observations and adjusts its reactions [Russel and Norvig (2021)].

There are three basic forms of machine learning.⁴ In supervised learning, the AI is programmed to map input to output [Russel and Norvig (2021)]. Input might be an image and output the classification as a wolf. The database that trains the AI contains labeled examples. The label tells the AI which function to find (hence the term “supervised” learning). Supervised learning requires large datasets that have been processed and appropriately labeled. Using these, the AI learns to make predictions for new data.

Some situations require a more exploratory approach. The goal might be to analyze unlabeled data with a clear goal in mind. Alternatively, it might not even be clear which questions are relevant; for example, when dealing with a large, unstructured dataset [Russel and Norvig (2021)]. Unsupervised learning responds to these exploratory needs. It makes the AI independently find structures and patterns. The programmer does not specify the way in which the AI performs the identification task, nor do they specify a goal or label the data. This distinguishes the technique from supervised learning, where the AI has a previously known objective. With unsupervised learning, the AI shows the user a way of sorting disordered data. This approach requires very large datasets and computers with enormous computing power. Its use for daily management will for most corporations mean buying the AI from a provider.

Learning by reinforcement occupies a space between supervised and unsupervised learning [Russel and Norvig (2021)]. The AI works without pre-labeled training data and is programmed to perform certain sequences, such as a board game [Russel and Norvig (2021)] or a robotics task [Ertel (2021)]. It receives positive or negative human feedback after completing its task. Each following round, the AI adapts its strategy to receive positive feedback [Russel and Norvig (2021)].

³ Exploratory data analysis precedes making inferences and producing testable hypotheses. It does not include formal statistical modeling and inference. Instead, it helps to see patterns in the data, catch mistakes, and generate potential hypotheses.

⁴ Russell and Norvig (2021), p. 671.

“

Fresh efforts must go into understanding what corporate law expects from board members who rely on AI support to augment their decision making.

”

2.2 Induction engines

To a statistician, it comes as no surprise that good data is a core ingredient for a forceful prediction. Selection biases, omitted variable biases, or the non-observance of confounding variables can be just as damaging as mathematical errors in a model. With AI, many of these issues arise in similar ways. Depending on which training data the AI receives, how that data is structured or labeled, the AI will learn to map, recognize patterns, and build models to assess future situations [Russel and Norvig (2021)]. Predictions based on a carefully curated [Data Governance Working Group of the Global Partnership of AI (2020)], possibly even synthetic [Jordan et al. (2022)], dataset differ significantly from the prognosis an AI makes by accessing the entire internet. If biases or past discrimination are baked into the data, the AI will suggest treating new cases in line with seasoned values. The same goes for the selection of data for the AI to learn [Russel and Norvig (2021)]. Consider the example of the bank executive deciding on branch closures. If the AI is trained on a small dataset, compiled by one bank, sampling customer reactions in one geographical area, the AI will develop a model that provides an excellent representation of this one dataset, but will not necessarily generalize. The risk of error increases and the quality of the prediction decreases.

This is not to say that more data is necessarily the better solution. Take open access to the internet as an illustration. It allows for particularly precise predictions about human preferences, detecting unanticipated patterns and clusters. At the same time, much of the data is noise that risks producing skewed results [Bender et al. (2021)]. To bolster management decisions, a synthetic, curated, or at least “cleaned” dataset might be more useful.

Lastly, it is helpful to keep in mind that AIs are “induction engines” [Larson (2021)]. Their probabilistic estimations rely on correlations that they infer from existing data. A change in circumstances, unusual, or rare, situations, technical innovations, or novel human preferences arrive at an AI with a time delay [Marcus (2018)].

3. DECISION MAKING AND CORPORATE LAW: OWNERSHIP AND TRUST

Decision making is one of the areas where AI has been shown to augment human capabilities. There are preformatted and rule-bound situations that provide especially fitting use cases for AI. We might be looking at robots for production, a chatbot used on a customer hotline, or automated lending decisions. Along similar lines, the AI might take over parts of rule-based decision making. Consider a chatbot forwarding unfamiliar questions or an out-of-the ordinary credit application that human employees review further. Board decisions, by contrast, are rarely an exclusively rule-based endeavor [March (1994)]. They entail discretion, intuition, and “gut”, a process of weighing and balancing different considerations, and of making value judgments. Employing AI as a prediction machine makes it possible to build scenarios, assess their probability of materializing, and use this as a background when making an informed decision.

Corporate law adapts rights and duties to the different types of decision makers. It treats board members differently from decision makers at officer and employee level. Firstly, the law expects board members to fully own their decisions. By way of illustration, see Delaware General Corporate Law § 141(a) that provides that a Delaware corporation is managed by or under the direction of the board of directors. In discharging their duties, they owe fiduciary duties of loyalty and care. Secondly, and as a flipside of ownership, the board allows for trust in board members. As long as they act loyally and carefully, the business judgment rule provides a generous liability regime. While board members must critically review material information, they are not required to work through any and all available information. As a second best, the law accepts what is “simply bad judgment” by board members,⁵ rather than encouraging judicial second-guessing.

⁵ Joy v. North 692 F.2d 880, 885 (2d Cir.), cert. denied, 460 U.S. 1051 (1983); Bainbridge (2020).

4. THE BOARD'S ROLE IN STRUCTURING DECISION MAKING BY OFFICERS AND EMPLOYEES

Some board resolutions are purely organizational in nature. They allow for and structure decision making by officers and employees of the corporation. Oversight duties remain with the board. Arguably, bedrock principles of corporate law are well suited to cope with these board decisions. A board must assess the value proposition of integrating AI. Gains in speed and accuracy must be balanced against the availability of an AI that is fit for the intended purpose. Relevant data and options to train personnel must be evaluated, error costs if things go wrong must be assessed.

Decisions of that type are not the focus of this paper, which deals with an AI enhancing board decision making. Still, three remarks are in order to hint at the relevant duties of care. The availability of an AI model that is fit for the intended purpose is an obvious first consideration. Some departments, such as trading, compliance, or risk management might be especially prone to using AI in the form of ML. Marketing and customer services might profit enormously from LLMs. In other cases, integrating AI might require a rewiring of the entire workflow [Agrawal et al. (2022)]. Balancing the potential gain against the probable costs is a business judgment for which the law grants boards considerable discretion. This includes the suitability of the selected product, extends to its ongoing control, and follow-up product monitoring. In most cases, corporations will purchase the AI from a third party. Selecting an appropriate provider and making sure the offer can be tuned to data that is relevant for the corporation is relevant for the board's choice of an AI. Over time, standard practices will develop, shaping the business judgment on why to choose one AI over another. The E.U. AI Act encourages certifications and provides guidelines. In what it terms "high-risk applications", it includes mandatory requirements that will shape board choices for an appropriate AI.⁶

If the choice of the AI model is the first step, the availability and relevance of data comes next. Business judgments concern questions such as: does the corporation have proprietary data or can it obtain third-party data at a reasonable cost? What type of data is needed (for instance, open source, curated, synthetic, labeled, etc.), how high is the probability of flawed data, and how high are estimated costs when proceeding

with it? Will the AI be helpful as a "cognitive fix" for standard flaws of human decision-making? How high is the risk of the AI learning from biased decisions [Langenbucher (2023a)]?

Additionally, the intended "workplace" for the AI might require specific features. Employees who cooperate with an AI often need special skills.⁷ This involves basic training, as required for every new machine or technology, to be able to correctly classify its mode of operation and risks. The risk of known shortcomings of AI – for example, problems with the coding of known knowledge [Marcus (2018)] or abductive conclusions [Larson (2021)] – must be balanced against an increase in efficiency.

5. THE BOARD STRUCTURING ITS OWN DECISION MAKING

Using AI as a prediction machine when preparing a board decision is different from the board adopting AI as a tool to enhance the corporation's workflow. Rather than programming (partly or fully) automated decisions, the board integrates the AI into its own deliberations. It hopes to enhance the quality of its decision making by gaining a good understanding of, for instance, how markets, customer preferences, capital allocation, or investor appetite will evolve.

Board resolutions of this type operate under the corporate law principles mentioned above [Langenbucher (2023c)]. On the one hand, the business judgment rule represents the law trusting board members with decision making and keeping judicial second-guessing to a minimum. On the other hand, the law expects the board to own its decisions, ruling out an abdication of authority or an overreliance on experts. The requirement to own a decision leaves no room for the board to have an AI decide in its place. At the same time, the law says nothing against the board asking for support in its decision making [Fleischer (2023)]. An emerging discussion has revolved around how to draw the line between an AI merely supporting and entirely taking over decision making. I argue below that (where we stand today) it is unlikely to see a board so comprehensively integrate an AI in its decisions that we would be looking at an abdication of board authority. Instead, I suggest that fresh efforts must go into understanding what corporate law expects from board members who rely on support to augment their decision making. I use Delaware corporate law to illustrate legal rules for human experts who assist the board and suggest adapting these to the challenges brought about by integrating AI into board decision making.

⁶ Annex III spells out the high-risk AI systems referred to in Art. 6(2) AI Act.

⁷ For "automation bias" see Art. 14(4)(b) AI Act.

5.1 Abdicating authority: Does AI take over?

Traditionally, abdication has been understood as trading away the board's discretion.⁸ Against that background, so-called "black-box" AI has troubled some scholars [Dubovitskaya and Buchholz (2023)]. They view integrating a black-box AI as an abdication of authority to an "AI-oracle", as it were. The problem with AI as a tool augmenting decision making, they claim, is especially prominent if its predictions and recommendations cannot be explained. This view is rejected here as focusing overly on one element of a decision, losing sight of the broader board judgment [Langenbucher (2023c)].

Many scenarios are straightforward. It does not hurt to prepare a board decision by googling relevant facts. Another clear case is the (more theoretical) scenario of a board that formally or effectively commits to follow an AI's recommendation. Arguably, the law will not treat this situation any different from a board that trades away its authority to a human [Fleischer (2023), Möslin (2018)]. The relevant issue at stake is the same to the extent that the board does not have discretion to decide as it seems fit. From this perspective, it does not matter whether the AI is explainable or not.

The hard cases are situated between these two scenarios. With AI developing into a standard tool, board judgments will look and feel differently than today. AI outperforms humans in many tasks and continues to evolve, taking over ever more areas. A clear distinction between the AI preparing the decision and the board making the decision will often look artificial [Langenbucher (2024), advocating for a distinction along those lines: Fleischer (2023) and Noack (2019)]. The more closely a decision follows the AI's recommendation, the more the board's role might seem reduced to implementing what the AI has proposed [Agrawal et al. (2022)]. Arguably, building a basic understanding of technology and trying to grasp the inherent logic of algorithms provides some relief [Fleischer (2023)]. Still, few board members will become experts in AI technology.

Additionally, it does not help that humans are known to be subject to a wide variety of decision making anomalies when it comes to assessing statistical probabilities [Kahneman (2012), Kay and King (2020), Tversky and Kahneman (1983), Tversky and Kahneman (1974), Kozyreva et al. (2019)], a core element of AI. In the same way that the AI's "workplace" on any of the corporation's hierarchical levels must be carefully structured, the board's own "workplace" in cooperation with an AI also needs structure. Human cognition follows different patterns

than an AI [Burton et al. (2020)]. This entails thinking about the appropriate cognitive cooperation with the AI. Sometimes, the AI can be very helpful if it acts as a "cognitive fix" for human behavioral anomalies [Burton et al. (2020)], however, the more behavioral anomalies have been baked into the data the AI was trained on, the more these are amplified at scale, rather than reduced. Additionally, scholars have highlighted human preferences for social interaction instead of receiving algorithmic advice [Burton et al. (2020)]. If offered the choice, humans seem to go for a discursive back and forth, rather than receiving a blunt prognosis without the option to engage in arguments and counter-arguments [Miller (2023)]. When the stakes are high, humans tend to demand "slow and effortful consideration of evidence", even if empirical evidence does not necessarily find that this strategy leads to better decision making [Burton et al. (2020)].

Against this background, the tough question corporate law must answer is what it expects as a minimum from board members in terms of owning decisions that rest on predictions by an AI. Arguably, the prohibition to abdicate board authority is too coarse a tool to provide a meaningful answer. While few board members have a precise understanding of how everyday AI, such as the Google search engine or ChatGPT, produces its results, the same is true for a pocket calculator or a GPS. The reason we do not consider the use of these tools as abdication of board authority to a machine is that they contribute but one element to a decision that the board fully owns. It follows from there that the relevant question is not if but how board members integrate AI in their overall judgment. Short of a situation where the board commits to following the AI's recommendation "no matter what", most cases are not about abdicating authority. Instead, they have to do with delegating (increasingly large) parts of the decision making process.

5.2 Informing board decision making: When to trust an AI

Most board decisions rest on a large variety of assumptions and predictions. Many of these are known unknowns: how will the market react to the bank closing brick-and-mortar branches? Will the self-driving car cause terrible accidents? Which percentage of my debtors will perform on their loan? When will customer preferences for my product change? How will a geopolitical crisis affect my firm? In these scenarios, the board owning its decision translates as: understanding the risk of working with a known unknown, evaluating it, and

⁸ See for Germany: Telle (2023) and Möslin (2018). For the U.S., see Bruner (2021) and Petrin (2019).

forming an informed and reasonable judgment. The prediction that an AI makes, explainable or black-box, can be just that: a known unknown.

So far, we have seen that the law allows boards to delegate individual parts of a decision making process. This includes a decision in the face of known unknowns. With these, the law trusts board members to come to a reasoned business judgment. Nonetheless, boards do not get a *carte blanche*. Generally, a board must evaluate and double-check the information it receives. On closer inspection, the law distinguishes between both decisions (business judgments and others) and support tools (technical help, humans integrated in the corporation, outside experts).

Board members' duties of care vary depending on the decision at hand. For business judgments, the law largely trusts the board, lowering its standard of judicial review. As to doctrinal detail, jurisdictions follow different approaches. Under Delaware law, it is for the plaintiff to prove that the board did not collect appropriate information before making a business judgment.

Outside of business judgments, courts apply an enhanced scrutiny standard. Compliance and risk management are paradigm examples. Courts assess the board's decision making process, including the information the board collected and evaluated.⁹ Sometimes, this can restrict the use of AI, especially of the black-box variant.

Consider a board that wishes to cut down on costs. It is impressed by an AI that performs better at predicting credit default risk of borrowers or suitability of potential new hires. It decides to restructure its human resources or its credit underwriting department. Many elements of this plan qualify as a business judgment – the need to cut down costs, the choice between different AI models, the decision to remodel the entire department, or start with small steps. However, some elements of the board's decision do not qualify as a business judgment with its ensuing broad discretion. The lively debate on algorithmic discrimination provides ample examples for such elements [Langenbacher (2023a)]: the decision to restructure human resources must not lead to hiring decisions that systematically discriminate between applicants. Automating credit underwriting must not allow

for discriminatory lending practices. Assume, as an integral part of restructuring credit underwriting, the board installs a black-box AI to help with assessing credit default risk. Anti-discrimination laws, such as the U.S. Equal Credit Opportunities Act or the E.U. Consumer Credit Protection Directive prohibit a denial of credit based on protected characteristics. Assume further that the AI collects publicly available data on retail consumers, develops personalized credit default risk assessments, recommends underwriting decisions, or even extends an automated contractual offer. To respect anti-discrimination law, the AI is programmed to disregard all protected attributes. However, given the big data it draws on, the AI is still likely to use proxy variables. Proxy variables stand in for protected characteristics. First names may double as gender or ethnicity, social media friends can be a proxy for age, and activities on a Saturday a proxy for religious faith. The use of proxy variables (first name) can lead to a disparate outcome between minority and majority groups (women and men), even if no protected characteristic (gender) was used. Neither the board nor the corporation's credit officers, or even data scientists and coders of the AI, will necessarily be able to identify the variables that the black-box AI used.

Can the board reason as follows:

- We understand that the board must not allow credit underwriting decisions to vary along a protected variable.
- Our AI is programmed to disregard protected variables when making its prediction.
- We understand that this AI might use proxy variables, but the extent to which it does is a known unknown.
- The law trusts boards to integrate known unknowns in its decision if the board evaluates the ensuing risk.
- As long as we assess the profit to be made with the credit-underwriting AI and balance it against the risk of potential litigation, we are fine to use the black-box AI.

Assuming an affirmative duty for the board to obey the law,¹⁰ the decision regarding whether we face a known-unknown scenario depends on an interpretation of the anti-discrimination rule. Courts might decide that compliance with that rule requires nothing more than the installation of an input restriction for protected characteristics.¹¹ Following

⁹ Delaware law adds major decisions, such as change of control transactions, the sale of the company, or the implementation of defenses in a takeover situation. Under German (and European) law, these are business judgments but require shareholder consent.

¹⁰ Proponents of (some version of) the efficient breach theory do not share this assumption, for a foundation see Posner (2009); for an overview see Bigoni et al. (2014).

¹¹ See the upcoming German law on credit scoring, German Federal Government (2024).

this interpretation, the black-box AI could be used, as long as the input restriction was in place. Courts that prefer a tougher reading of the anti-discrimination rule might introduce further restrictions on permitted data,¹² or prohibit the use of black-box AI altogether. What distinguishes this scenario from the AlphaFold¹³ example is the degree of trust accorded to the board. The decision to restructure credit underwriting as such is up to the board. However, the decision to install a black-box AI to hand out loan contracts is not entirely discretionary. As far as protected groups are concerned, the law requires some degree of scrutiny as to the known unknown element. This stands in contrast with the AlphaFold scenario. The board was able to treat AlphaFold and its findings on protein folding structures as a known unknown, qualifying as a classic business judgment.

When informing the board, technical support tools, ranging from a pocket calculator to high-powered computers, have been a standard feature. There are no rules stipulating distinct duties for boards that employ a machine to assist decision making. This is different from situations where humans support board decision making, especially if the human help is not an employee of the corporation. The law expects some level of engagement from a board that has humans inform its decision making. DGCL § 141(e) distinguishes between “information, opinions, reports, or statements presented by any of the corporation’s officers, employees, or committees” and input “by any other person.” A board may draw on sources from inside the company as long as this is done in good faith. For outside experts, the rule adds extra test prongs. The input must stem from “any other person as to matters the member reasonably believes are within such other person’s professional or expert competence.” Additionally, such person must have “been selected with reasonable care by or on behalf of the corporation.” Hence, for outside experts the court will explore two issues: the reasonable belief that the expert is competent to deliver the relevant input, and the reasonableness of the director’s selection of the expert. For both issues, the standard of review is strict, and the business judgment rule is not available.¹⁴

5.3 Visualizing ownership and control

To illustrate how courts review board decision making, I have provided a four-square control matrix in Figure 1. The y-axis represents the level of allowance for board discretion

according to the decision’s subject matter (trust). Boards enjoy broad discretion for those elements of a resolution that qualify as a business judgment. Little discretion is accorded to parts of a decision that have to do with compliance, risk management, and similar non-business-judgment issues. The x-axis looks at the intensity of information support (ownership). Boards have been free to use technical support tools, ranging from pocket calculators to high-powered computer networks. Human helpers have attracted more scrutiny. This is true for input by officers, committees, or employees of the corporation. Even more scrutiny concerns outside experts.

Following this representation, four squares emerge. In the upper right-hand corner we find the first square, which symbolizes business judgments that score high on discretion and do not rely on human support. They face the lowest degree of judicial review.

The second square is situated in the upper left-hand corner, and symbolizes decisions that still score high on discretion but have drawn on considerable help, including from outside the corporation. For those, the standard of review is higher than for the first square, given the dominant role of support tools.

A third square is situated in the lower right-hand corner. It symbolizes decisions that are about non-business judgment issues but do not rely much on human support. Its standard of review resembles the one just described. It is higher than for the first square, given its low score on trust.

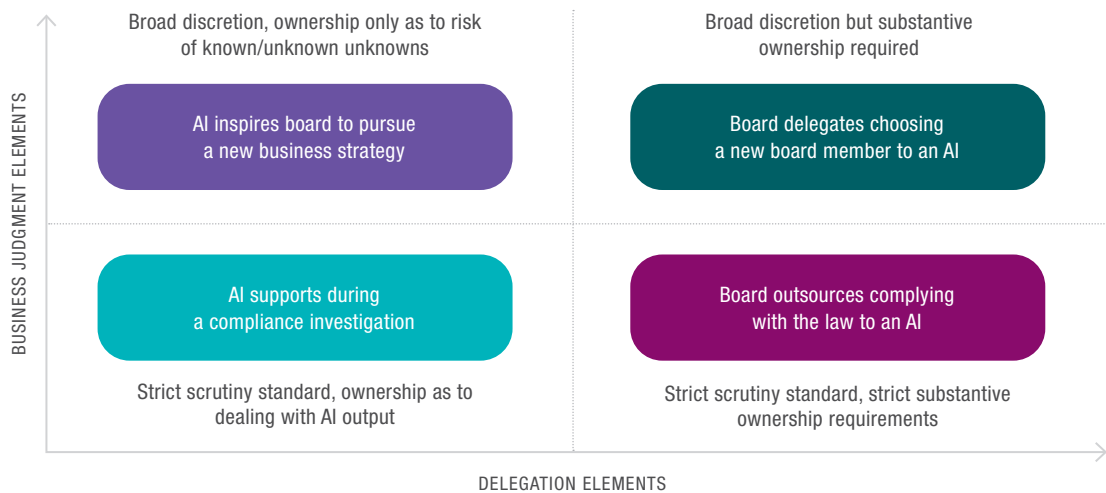
The fourth square, which is located in the lower left-hand corner, shows decisions that were reached with much outside help, hence, score low on ownership. Additionally, these decisions score low on discretion, because they include few or no business judgment elements. This square symbolizes the highest intensity of judicial review.

The graphical representation is helpful given that board decisions rarely fall into one neat category. The above example on restructuring credit underwriting showed how board decisions combine different elements. Some of these are about developing and deciding on a novel business strategy, involving market knowledge, experience, intuition, and gut. All these are characteristics of a low-judicial-scrutiny decision. However, other parts of the decision might depend on the professional evaluation of a particular market niche or of a new product that only outside experts can deliver. Legal issues

¹² See Art. 18(3) Consumer Credit Directive (EU) 2023/2225 on data gathered from social media.

¹³ AlphaFold is an artificial intelligence program developed by DeepMind, a subsidiary of Alphabet, which performs predictions of protein structure. The program is designed as a deep learning system (<https://tinyurl.com/wmvkjfha>).

¹⁴ *Brehm v. Eisner*, 746 A.2d 244, 262 (Del. 2000), <https://tinyurl.com/ywmswc9c>

Figure 1: How courts review board decision making

might be decisive for the success of the new strategy because a new product requires regulatory approval. These legal issues could be small and resolvable in-house or complex, calling for outside counsel. Visualizing the matrix and “moving” the decision, as it were, allows us to understand the degree of judicial review that a comprehensive board resolution, with its various sub-parts, will attract.

The legal logic underlying the matrix reflects the tension between boards owning their decisions and the law trusting boards without holding them accountable for “simply bad judgment”. As explained above, the law expects that board members are accountable and will own their decisions. It follows from there that the law does not allow the board to abdicate its authority and hide behind an alternative decision-maker, as it were. It does not matter whether an alternative decision maker might be more capable than the board: it is not the one the shareholders voted for. Along similar lines, the board may not delegate core parts of its decision making to non-board members. The more a board decision looks like nodding to what someone else has proposed, the less it conforms with the law’s expectation of the board owning its decision.

At the same time, a board cannot sensibly own a decision unless it fully understands its pros and cons. If the board lacks the relevant knowledge or if it would take too much time to gain comprehensive insight, it makes sense to bring in help. However, human helpers come with their own sets of thoughts, approaches, and incentives that are not necessarily transparent to the board. Additionally, the board members might lack the expert knowledge to evaluate their input.

Delaware law is mindful of that, distinguishing between the type of human helpers a board brings in. If these are officers or employees of the corporation, the trust the law places on boards by and large extends to these helping hands. With outside experts, it is less clear that their incentives are aligned with the corporation in the way officers and corporate committees are. Against this background, Delaware law allows the board to trust outside experts but tightens the requirements for doing so, by stressing the careful selection of the expert, including their field of expertise.

A board that uses an AI prediction as its stepping stone is likely to face liability if an overreliance on the flawed AI-prediction led to a bad business decision. Following the control matrix visualization, a first line of defense shows. Corporate law trusts board members to exercise discretion whenever a business judgment is at stake. Substantive control of what the board considered the best business strategy is low because the law is reluctant to make judges second-guess managerial decisions.

However, the trust placed on board members comes with the expectation that they own their decisions. This points towards the second line of defense. A board that painstakingly double-checks the information it receives fully owns its decision. By contrast, the more a board outsources important parts of decision making to inside or outside help, the stricter the judicial review, the more intense the relevant duties of care for selecting help. By way of illustration: Delaware judges will double-check the board’s selection of an expert.

The visualization of the control matrix shows how it is neither necessary to comprehensively define any AI as a purely technical support tool, nor to unflinchingly analogize an AI to a human expert, be it inside or outside the corporation. Instead, the matrix allows to move the needle, as it were, along the x-axis, ranging from low to high ownership. The everyday AI search engine resembles the purely technical support tool that corporate law has not deemed to be in need of special judicial scrutiny. This is true for both business judgments and non-business judgments. The same can be true for a very sophisticated AI that inspires the board to move ahead with a novel product. Its decision concerns a business judgment that the law entrusts to the board. Deciding in the face of a known unknown along those lines is anything but unusual for a corporate board. Putting a probability on different outcomes and deciding which risk to take when faced with uncertainty is what the law trusts the board to do. Visualizing it in the control matrix, we look at the upper right-hand corner.

Using an AI in a credit underwriting scenario is a counter example. Assume an AI furnishes an assessment of credit default risk. One element of the decision to restructure the credit underwriting department concerns pricing loans, a standard business judgment that qualifies for a high level of trust towards the board. However, a major part of credit underwriting has to do with compliance with anti-discrimination laws. For those parts, there is low discretion accorded to the board. The board is not faced with a known-unknowns situation. It is not the board's task to put a probability on its credit model breaking the law and then move forward, in line with its risk appetite. Instead, we face a scenario where strict substantive control is in order. For a board to fully own a decision about complying with the law, it must make sure it has gathered enough information to not break the law. Visualizing the y-axis of the matrix helps to identify the level of judicial scrutiny. A black-box model that produces automated underwriting decisions achieves a very low score of ownership and, in turn, makes a case for intense judicial scrutiny. By contrast, an explainable model, working exclusively with a limited list of known data points, scores high on ownership. It makes it possible to assess individual credit underwriting decisions. The board might not be able to converse with the AI like it would with a human peer, but it has access to an explanation regarding why the AI preferred one loan over another.

6. CONCLUSION

This article has explored the legal ramifications of board members employing AI to augment their decision making. It focuses on AI as "prediction machines" that offer a glance into

the future. I submit that predictions, with or without AI, are an everyday element of board decision making. They imply an assessment and a risk evaluation of known unknowns, a paradigmatic example for a business judgment. Corporate law is well aware of the necessity to trust boards with making such decisions. Still, the law requires board members to eventually own their decisions, rather than diffuse responsibility among the various helpers that inform boards.

Two dimensions, ownership and trust, provide the framework for understanding how corporate law shapes board decision making. I introduce a "control matrix" to graphically illustrate these dimensions. If the law accords high levels of trust to the board, we look at business judgments that offer considerable discretion. Low levels of trust are characteristic for rule-bound decisions such as compliance. High levels of ownership characterize decisions that the board takes, by and large, without external support. The more elements of a decision a board outsources to officers, committees, or outside experts, the lower its ownership of the final board decision.

Augmenting decision making via an AI, I claim, does not necessarily amount to a loss of ownership. Importantly, it does not involve a novel form of abdicating board authority. This applies to both explainable and black-box AI. Rather, using an AI to inform boards can be understood in the broader context of boards drawing on support in the form of technical tools or internal and external experts.

To fully understand the relevant standard of judicial review, the dimension of ownership must be complemented by its twin dimension of trust. I introduce a graphic representation to allow for situating a board decision along these two dimensions. Business judgments score high on trust. This makes for a flexible standard of judicial review. By contrast, non-business judgments fall under an enhanced standard of judicial review.

A board that comprehensively builds a non-business judgment on an AI prediction scores low on both dimensions, ownership and trust. It faces intense judicial review. By contrast, a board that uses AI merely to inspire a classic business judgment scores high on both dimensions, entailing low judicial review. Two scenarios sit in between. A business judgment that relies predominantly on an AI prediction scores high on trust but low on ownership and a non-business judgment that the board takes with little help from an AI scores low on trust but high on ownership. Building on this framework, future research endeavors will have to spell out the details of relevant duties of care.

REFERENCES

- Agrawal, A., J. Gans, and A. Goldfarb, 2018, *Prediction machines: the simple economics of artificial intelligence*, Harvard Business Review Press
- Agrawal, A., J. Gans, and A. Goldfarb, 2022, *Power and prediction: the disruptive economics of artificial intelligence*, Harvard Business Review Press, Boston
- Bainbridge, S. M., 2020, *Corporate law*, 4th revised edition, Foundation Press
- Bender, E., T. Gebru, A. McMillan-Major, and S. Shmitchell, 2021, "On the danger of stochastic parrots: can language models be too big?" *Proceedings of FAccT '21*, 610–623
- Bigoni, M., S. Bortolotti, F. Parisi, and A. Porat, 2014, "Unbundling efficient breach," *University of Chicago Coase-Sandor Institute for Law & Economics research paper no. 695*, <https://tinyurl.com/s9mmb4m9>
- Bruner, C. M., 2021, "Artificially intelligent boards and the future of Delaware Corporate Law," *University of Georgia School of Law, Legal Studies research paper no. 2021-23*, <https://tinyurl.com/ys7p5m3d>
- Burton, J. W. M. K. Stein, and T. B. Jensen, 2020, "A systematic review of algorithm aversion in augmented decision making," *Journal of Behavioral Decision Making* 33, 220-239
- Data Governance Working Group of the Global Partnership of AI, 2020, "The role of data in AI," *Report for the Data Governance Working Group of the Global Partnership of AI*, <https://tinyurl.com/273rpdht>
- Dubovitskaya, E., and A. Buchholz, 2023, "Die Geschäftsleitung und der Rat des Algorithmus," *ZIP* 2023, 63-73
- Ertel, W., 2021, *Grundkurs Künstliche Intelligenz: Eine praxisorientierte Einführung*, Springer
- Fleischer, H., 2020, *Beck-online Großkommentar German Federal Government*, 2024, *Entwurf eines Ersten Gesetzes zur Änderung des Bundesdatenschutzgesetzes*, <https://tinyurl.com/48v25m42>
- Glimcher, P., 2004, *Decisions, uncertainty, and the brain: the science of neuroeconomics*, The MIT Press
- Hawkins, J., 2021, *A thousand brains. A new theory of intelligence*, Basic Books
- Jordan, J., F. Houssiau, G. Cherubin, S. N. Cohen, L. Szpruch, M. Bottarelli, C. Maple, and A. Weller, 2022, "Synthetic data – what, why and how?" *The Royal Society*, <https://tinyurl.com/52nuxsby>
- Kahneman, D., 2012, *Thinking, fast and slow*, Penguin Books
- Kay, J., and M. R. King, 2020, *Uncertainty, decision-making for an unknowable future*, The Bridge Street Press
- Koch, J., 2023, in *Aktiengesetz*, 17th edition, Beck
- Kozyreva, A., T. Pleskac, T. Pachur, and R. Hertwig, 2019, "Interpreting uncertainty: a brief history of not knowing," in Hertwig, R., T. Pleskac, and T. Pachur (eds.), *Taming uncertainty*, MIT Press
- Langenbacher, K., 2023a, "Consumer credit in the age of AI – beyond anti-discrimination law," *ECGI working paper series in Law no. 663/2022*, <https://tinyurl.com/ye8mh3nd>
- Langenbacher, K., 2023b, "KI in der Leitungsentscheidung des Vorstands der Aktiengesellschaft," *SAFE white paper no. 96*, <https://tinyurl.com/4s3yd3rf>
- Langenbacher, K., 2023c, "Künstliche Intelligenz in der Leitungsentscheidung des Vorstands," *ZHR* 187, 723-738
- Langenbacher, K., 2024, "AI judgment rule(s)," *forthcoming, Chicago Law Review online*
- Larson, E. J., 2021, *The myth of artificial intelligence. Why computers can't think the way we do*, Harvard University Press
- March, J. G., 1994, *A primer on decision making. How decisions happen*, The Free Press
- Marcus, G., 2018, "Deep learning: a critical appraisal," <https://tinyurl.com/z99rczsn>
- Miller, T., 2023, "Explainable AI is dead, long live explainable AI!. Hypothesis-driven decision support," <https://tinyurl.com/mr275aut>
- Möslein, F., 2018, "Robots in the boardroom: artificial intelligence and corporate law," in Barfield, W., and U. Pagallo (eds.), *Research handbook on the law of artificial intelligence*, Edward Elgar Publishing
- Nath, R., 2009, *Philosophy of artificial intelligence: a critique of the mechanistic theory of mind*, Universal Publishers
- Noack, U., 2019, "Organisationspflichten und -strukturen kraft Digitalisierung," *ZHR* 183, 105-144
- Petrin, M., 2019, "Corporate management in the age of AI," *Columbia Business Law Review* 3, 965-1030
- Posner, R., 2009, "Let us never blame a contract breaker," *Michigan Law Review* 107, 1349-1364
- Rolffs, M., 2023, "Causality and mental causation, a defense of nonreductive physicalism," *Stanford Encyclopedia of Philosophy*, <https://tinyurl.com/4ca94f65>
- Russell, S., and P. Norvig, 2021, *Artificial intelligence: a modern approach*, 4th edition, Pearson
- Stiehl, T., and A. Marciniak-Czochra, 2021, "Intelligente Algorithmen und Gleichungen? – Eine Annäherung an die Intelligenz mathematischer Konzepte," in Holm Hadulla, R. M., J. Funke, and M. Wink, (ed.), *Intelligence: theoretical foundations and practical applications*, Heidelbergberger Jahrbücher vol. 6
- Spindler, G., 2023, *Münchener Kommentar zum Aktiengesetz*, 6th edition, Beck
- Telle, M., 2023, "Einsatz Künstlicher Intelligenz zur vorbereitenden Unterstützung von Leitungsentscheidungen des Vorstands einer AG," in *Abhandlungen zum Deutschen und Europäischen Gesellschafts- und Kapitalmarktrecht (AGK)*, Volume 209
- Tversky, A., and D. Kahneman, 1974, "Judgment under uncertainty: heuristics and biases," *Science, New Series* 185:4157, 1124-1131
- Tversky, A., and D. Kahneman, 1983, "Extensional versus intuitive reasoning: the conjunction fallacy in probability judgement," *Psychological Review* Volume 90:4, 293-315

THE TRANSFORMATIVE POWER OF AI IN THE LEGAL SECTOR: BALANCING INNOVATION, STRATEGY, AND HUMAN SKILLS

EUGENIA NAVARRO | Lecturer and Director of the Legal Operations and Legal Tech Course, ESADE

ABSTRACT

The integration of artificial intelligence (AI) in the legal sector presents significant opportunities for improving efficiency, automating repetitive tasks, and enhancing decision making processes. However, successful implementation requires a clear strategy, proper training for legal teams, and the right collaboration between internal and external experts. Generative AI (GenAI) can streamline document drafting and client interaction, while non-generative AI excels in predictive analytics and e-discovery. Despite the advancements, AI cannot replace human emotional intelligence, creativity, and ethical judgment, which remain critical in delivering personalized and high-quality legal services. Ultimately, AI is a powerful tool, but its true value lies in complementing human expertise, not replacing it.

1. INTRODUCTION

Many lawyers and firms see GenAI as a sort of “magic button” that will automatically solve problems, without understanding that the real value that comes from identifying a clear use case. Instead of asking “how do we implement GenAI?”, the question should be “what needs do we have that GenAI could help solve?”

Some of the common mistakes that many make include:

- **Lack of a clear strategy:** GenAI is implemented without a specific objective, which can lead to investments in tools that are not suited to the existing workflow.
- **Unrealistic expectations:** lawyers sometimes believe that AI can completely replace human work, when in reality its greatest value lies in complementing human capabilities, automating repetitive tasks, or analyzing large volumes of data.
- **Resistance to change:** many legal professionals see technology as a threat rather than an opportunity to improve their practice. This can delay or complicate the effective implementation of any technological solution.

The key to successfully implementing GenAI, or any technology, is not to see it as a final solution, but as a tool that facilitates true innovation. The focus should be on improving the way lawyers work, collaborate, and deliver value to their clients.

The firms that achieve successful integration of technology are those that adopt a strategic and thoughtful approach. They do not just buy technological tools because it is a trend, but because they have clearly identified a problem that needs solving.

2. HOW TO CLEARLY DEFINE THE NEEDS THAT GenAI IS TO ADDRESS AND MANAGE THE IMPLEMENTATION CHALLENGE

For a firm or legal department to take full advantage of GenAI, it is essential to clearly define what need they seek to address. This can include:

- **Automating repetitive tasks:** such as creating standard legal documents, reviewing contracts, or drafting reports.
- **Analyzing large volumes of information:** using AI to process and summarize past case information, helping lawyers quickly access the most relevant precedents.

- **Improving client/business partner service:** through chatbots or virtual assistants that can answer basic client questions and guide them through their processes.

Each of these examples requires a detailed assessment of current workflows, the areas where lawyers are spending the most time on repetitive or manual tasks, and the points where automation or content generation could provide real value.

When it comes to implementing an AI project, choosing the right internal and external teams is crucial. AI projects, particularly in complex industries like the legal sector, require a diverse set of skills, expertise, and strategic alignment to be successful. The choice of teams not only impacts the quality of the implementation but also determines how seamlessly the new technology will be integrated into the organization.

2.1 Choosing the right internal team

The internal team plays a key role because they are the ones who understand the specific needs, workflows, and pain points of the organization. It is essential to select team members who are not only technically skilled but also deeply familiar with the company's operations and long-term objectives. Lawyers, IT staff, and project managers must collaborate closely to ensure that the AI solution addresses real business challenges.

Moreover, internal stakeholders need to be champions of the AI project, facilitating its adoption and supporting the necessary change management within the organization. Having a team that is committed, flexible, and open to learning new tools is essential for a smooth transition.

In some cases, the legal team is made up of traditional lawyers who may not be equipped to fully leverage the opportunities that a technological project offers, making it necessary to provide them with training. Training in technology, change management, and project management is crucial for the success of any technological initiative.

2.2 The role of the external team

The external team – whether it is an AI consultant, vendor, or development partner – brings in specialized expertise that the internal team may lack. However, the right external team should not only have technical knowledge; they should also understand the unique context of the legal or business environment they are working in. It is important to partner with experts who have experience implementing AI solutions in similar industries and can provide insights into best practices, potential challenges, and effective strategies. A key recommendation would be to ensure that the external team

“

We don't aim to be leaders in implementing technology or AI itself, but rather in finding real use cases that drive value to the company's strategy.

”

has experience in implementing projects specifically within the legal sector, as this greatly facilitates the process.

Equally important is ensuring that the external team aligns with the organization's vision and goals. Their role is not just to deliver a technology solution, but to act as a strategic partner, helping to guide the AI implementation in a way that maximizes business value.

2.3 The need for a collaborative environment

For an AI project to be successful, creating a “collaborative environment” is just as important as selecting the right teams. Collaboration fosters communication and transparency, which is key to aligning the internal and external teams. When both teams work together seamlessly, they can better anticipate challenges, address concerns, and adapt the project as needed.

A collaborative environment encourages innovation and problem solving. Internal teams provide real-world insights about the organization's needs, while external experts offer technical solutions that can be customized and refined. This iterative process is what ensures that the AI project is not just technically sound, but also practical and beneficial for the business.

Additionally, collaboration helps to create a sense of shared ownership over the project. When both internal and external teams are invested in the outcome, there is greater accountability and motivation to see the project succeed.

Choosing the right internal and external teams for an AI project is a critical step that can make or break the implementation process. The internal team brings the necessary knowledge of the business, while the external team provides the technical expertise and guidance required to successfully deploy AI solutions. Creating a collaborative environment between these teams ensures that both technical and strategic needs are

met, leading to a smoother integration of AI and a greater likelihood of long-term success. Ultimately, collaboration is the key to unlocking the full potential of any AI project.

3. WHY CORPORATE LAWYERS MUST LEAD TECHNOLOGY PROJECTS AND DEVELOPING SKILLS IN PROJECT MANAGEMENT

In today's rapidly evolving business landscape, the role of corporate lawyers is expanding beyond traditional legal advisory. With technology playing an increasingly central role in all areas of business, it has become crucial for corporate lawyers to step up and lead technological projects within their organizations. Embracing this role not only positions them as strategic partners but also enhances their ability to contribute to the company's growth and innovation.

3.1 Why corporate lawyers should lead technology projects

Corporate lawyers are uniquely positioned to lead technology projects because they have a deep understanding of the regulatory, compliance, and risk management aspects of the business. They are well-versed in the legal frameworks that govern technological advancements, such as data privacy, cybersecurity, and intellectual property rights, all of which are critical when implementing new technologies.

By taking the lead in these projects, lawyers can ensure that legal considerations are embedded in the design and execution from the outset, rather than being addressed as an afterthought. This proactive approach can prevent potential legal issues down the line and streamline the integration of technology into business operations.

Moreover, corporate lawyers bring a holistic perspective to technological projects, balancing legal risk with business opportunity. Their involvement can help align the project's goals with the broader strategic objectives of the organization, ensuring that technology is not just an operational tool but a driver of innovation and competitive advantage.

3.2 The need for corporate lawyers to develop project management skills

For corporate lawyers to successfully lead technological projects, it is essential that they also develop project management skills. Technological initiatives require careful planning, coordination, and execution across multiple departments, and lawyers need to be equipped to handle the complexities of these projects. Project management provides

the framework for setting clear goals, managing resources, and tracking progress. It ensures that deadlines are met, budgets are adhered to, and risks are mitigated throughout the lifecycle of the project. For lawyers, who are accustomed to working within strict legal timelines and managing complex deals, project management skills are a natural extension of their existing competencies.

Some of the important benefits of project management training for lawyers include:

- **Improved collaboration:** lawyers leading tech projects will need to collaborate with IT teams, external vendors, and other business units. Understanding project management methodologies helps facilitate communication and ensures all stakeholders are aligned and working toward the same objectives.
- **Risk management:** one of the core competencies of lawyers is identifying and mitigating risk. Project management allows lawyers to apply this skill in a structured way, identifying potential obstacles early and implementing strategies to mitigate them without delaying the project.
- **Efficiency and productivity:** legal departments are often seen as cost centers, but by taking a project management approach, corporate lawyers can lead projects that demonstrate value by improving operational efficiency and reducing costs through well-managed technological solutions.
- **Leadership and influence:** by leading technological initiatives with a strong project management approach, corporate lawyers position themselves as strategic leaders within the organization. This not only elevates their role but also enhances their influence across various business functions.

For corporate lawyers, taking the lead on technological projects is not just an opportunity, it is a necessity in today's digital age. By driving these initiatives, lawyers can ensure that legal compliance is built into the fabric of technological innovation. However, to do so effectively, it is equally important that they develop strong project management skills. This combination of legal expertise and project management acumen will enable lawyers to successfully navigate the complexities of technology integration, ensuring that projects are completed on time, within budget, and in alignment with the company's strategic goals. Ultimately, this shift empowers lawyers and to become true innovators and leaders in the business world.

4. THE NEED TO HAVE AN AI PROJECT

Many law firms and legal departments are mandated to have an AI project simply because it is trendy, but some have not optimized their processes or possess a document management system or a CLM, and most do not distinguish between AI and GenAI.

4.1 Understanding generative and non-generative AI

AI has emerged as a transformative technology across numerous industries, and the legal sector is no exception. Among the many types of AI, two categories have garnered significant attention: generative AI (GenAI) and non-generative AI. These categories differ in their underlying technologies and applications but share the common goal of improving efficiency and accuracy in legal work.

AI can be broadly divided into two categories: generative AI and non-generative AI. Understanding the difference between these two types of AI is crucial for their effective application in the legal sector.

4.1.1 UNDERSTANDING GenAI

GenAI refers to a class of artificial intelligence algorithms that can generate new content, such as text, images, code, or audio, based on patterns and examples from existing data. GenAI systems are typically based on models like Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), or large language models like GPT (Generative Pre-trained Transformers), which can create coherent content from a minimal set of inputs.

Key characteristics of GenAI include:

- **Content creation:** GenAI creates new data rather than just processing or analyzing existing data.
- **Self-improvement:** as it generates content, these systems refine themselves by learning from feedback loops, improving the quality and coherence of their outputs.
- **Examples in practice:** chatbots, document generation, automated legal advice, and case law synthesis are prime applications.

4.1.2 UNDERSTANDING NON-GENERATIVE AI

In contrast, non-generative AI is designed to process and analyze data without creating new content. It excels at identifying patterns, making predictions, and offering recommendations based on preexisting information. This type of AI includes machine learning models for classification, regression, clustering, and decision trees. Non-generative AI typically powers predictive analytics, pattern recognition, and decision support systems in various industries, including legal.

Key characteristics of non-generative AI include:

- **Analysis and prediction:** non-generative AI is geared toward analyzing data and making recommendations, such as predicting legal outcomes based on historical data.
- **Efficiency:** it optimizes workflows by processing large datasets and pinpointing insights that human practitioners might overlook.
- **Examples in practice:** legal research tools, e-discovery, document classification, and case outcome prediction.

5. THE ROLE OF AI IN THE LEGAL SECTOR

The legal sector, which has traditionally been slow to adopt new technologies, is now undergoing a paradigm shift due to the proliferation of AI tools. Both generative and non-generative AI are enhancing legal work, enabling faster decision making, automating repetitive tasks, and improving access to legal resources.

5.1 Applications of GenAI in the legal sector

5.1.1 AUTOMATED DOCUMENT DRAFTING AND REVIEW

One of the most significant contributions of GenAI in the legal field is its ability to automate document drafting and review processes. Law firms often spend considerable time drafting contracts, legal opinions, and briefs. With AI-powered document generation tools, attorneys can now input key data points or instructions, and the system generates a draft of the required document, saving time and reducing errors. For example, GenAI models like GPT can craft coherent and legally sound clauses in contracts by analyzing previous agreements. These systems can produce drafts with minimal human intervention, which is especially useful for tasks like non-disclosure agreements (NDAs), employment contracts, and merger and acquisition documents.

5.1.2 CHATBOTS FOR LEGAL ADVICE

Another area where GenAI is making a mark is through intelligent chatbots capable of providing basic legal advice. These chatbots use natural language processing (NLP) to understand user queries and generate relevant responses based on legal information databases. While such chatbots cannot replace human lawyers, they can handle preliminary questions related to legal procedures, assist clients in filling out forms, or provide initial guidance on various legal issues such as property disputes, family law, or small claims. This functionality helps reduce the workload for legal practitioners by addressing simple client queries and directing more complex matters to professionals.

5.1.3 CASE LAW SUMMARIZATION

GenAI has revolutionized legal research by providing tools capable of summarizing vast amounts of case law. These tools analyze legal precedents, generate concise summaries, and offer recommendations, enabling lawyers to quickly grasp the crux of a case. For example, AI-powered summarization tools can provide key insights from hundreds of legal documents in a fraction of the time it would take a human researcher. This capability is invaluable in cases with extensive case law or when preparing for court proceedings, as lawyers can identify relevant legal arguments and precedents more efficiently.

5.2 Applications of non-generative AI in the legal sector

5.2.1 PREDICTIVE ANALYTICS FOR LEGAL OUTCOMES

One of the most exciting applications of non-generative AI is its use in predictive analytics to forecast the outcomes of legal cases. By analyzing historical data, including past rulings, judge behaviors, and case specifics, AI models can predict the likelihood of a case's success. This helps lawyers assess the risks associated with pursuing litigation or advise clients on the best course of action. For example, tools like Lex Machina utilize machine learning algorithms to predict outcomes in intellectual property disputes, labor law cases, and other areas. Such predictive analytics offer valuable insights that can influence legal strategy and improve the chances of a favorable outcome.

5.2.2 E-DISCOVERY AND DOCUMENT CLASSIFICATION

In the discovery phase of litigation, legal teams are tasked with reviewing massive volumes of documents to identify relevant information. Non-generative AI plays a crucial role here by

streamlining the e-discovery process. AI-powered systems can analyze and categorize documents based on relevance, flagging key pieces of information for further review. These AI systems employ techniques like natural language processing (NLP) and machine learning to identify patterns and context within documents, ensuring that important information is not missed. By automating document review, AI allows lawyers to focus on high-level legal work while reducing costs and time spent on manual document searches.

5.2.3 FRAUD DETECTION AND COMPLIANCE MONITORING

Non-generative AI also excels in the field of fraud detection and regulatory compliance. By analyzing transaction patterns, communications, and contracts, AI systems can identify suspicious activities and flag potential legal or ethical violations. In areas like corporate law, financial regulation, and compliance, AI helps firms stay ahead of potential risks by offering real-time monitoring and alerts. Compliance systems powered by AI can scan emails, reports, and contracts to ensure adherence to regulations, helping businesses avoid costly legal penalties.

5.3 Challenges and ethical considerations

While both generative and non-generative AI offer considerable benefits to the legal sector, their adoption comes with challenges and ethical considerations. Legal professionals must navigate issues such as data privacy, transparency, and the potential for bias in AI algorithms.

5.3.1 DATA PRIVACY AND CONFIDENTIALITY

Law firms handle sensitive information that must be kept confidential. AI systems often require large datasets to function effectively, raising concerns about how client data is stored, processed, and protected. Legal professionals must ensure that AI tools comply with data privacy regulations like GDPR or HIPAA (Health Insurance Portability and Accountability Act), especially when dealing with client information.

5.3.2 TRANSPARENCY AND EXPLAINABILITY

Another challenge is ensuring transparency in AI decision making processes. In the legal sector, where decisions can have far-reaching consequences, understanding how AI arrives at its conclusions is critical. AI systems, particularly those using machine learning, are often referred to as "black boxes", meaning their decision making processes are opaque. This raises concerns about fairness and accountability, especially in cases where AI-generated recommendations are used in courtrooms.

5.3.3 BIAS AND FAIRNESS

AI models are only as good as the data they are trained on. If the historical data used to train these models contains biases, such biases can be perpetuated in AI-generated decisions. For example, if past judicial rulings reflect bias based on race, gender, or socioeconomic status, AI systems could reinforce these prejudices. Addressing this requires a concerted effort to audit AI models regularly and implement safeguards to ensure fairness.

6. THE FUTURE OF AI IN THE LEGAL SECTOR

Looking ahead, the role of AI in the legal sector will only expand. As AI tools become more sophisticated and integrated into legal workflows, they will further enhance the ability of legal professionals to deliver services efficiently. However, the human element remains critical – AI cannot fully replace the expertise, empathy, and ethical judgment of human lawyers.

GenAI will continue to evolve, particularly in drafting complex legal documents and conducting preliminary legal research. Non-generative AI will be essential in areas like predictive analytics and regulatory compliance, helping firms navigate an increasingly complex legal landscape.

Moreover, AI has the potential to democratize access to legal services by providing cost-effective tools for those who may not have had access to legal representation. Chatbots, document automation, and AI-driven legal advice platforms can serve as an entry point for individuals and small businesses seeking legal guidance.

Generative and non-generative AI are reshaping the legal sector, offering unprecedented opportunities for efficiency, accuracy, and innovation. While challenges like data privacy, transparency, and bias must be addressed, the potential for AI to transform legal work is undeniable. As these technologies advance, legal professionals who embrace AI will be better equipped to meet the demands of a rapidly changing industry.

Ultimately, the integration of AI in law is not about replacing lawyers but empowering them to deliver better, faster, and more equitable legal services.

7. CONCLUSION

AI has the potential to transform industries by improving efficiency and leveling the playing field in terms of access to knowledge. In sectors like law, AI can quickly process vast amounts of information, automate routine tasks, and provide valuable insights, allowing professionals to work faster and with greater accuracy. It can help ensure that even those with limited resources can access high-quality information and services, democratizing knowledge in ways that were previously unimaginable.

However, while AI excels at data processing and automation, it cannot replicate the emotional and interpersonal skills that are uniquely human. Empathy, creativity, critical thinking, and the ability to navigate complex human interactions are qualities that AI lacks. These emotional skills are increasingly becoming the key differentiators in professions where human connection and understanding are essential.

In areas such as client relations, negotiation, and conflict resolution, the ability to understand emotions, manage relationships, and adapt to changing social dynamics remains irreplaceable. Lawyers, for example, do not just need knowledge of the law; they need the emotional intelligence to listen to clients, understand their concerns, and provide guidance that takes both legal and personal factors into account.

As AI continues to handle more of the technical and repetitive aspects of work, the human element will stand out even more. Professionals who cultivate emotional intelligence will be better equipped to thrive in an AI-enhanced world, as their interpersonal skills will allow them to build stronger relationships, create more personalized experiences, and offer solutions that go beyond data-driven insights.

Ultimately, while AI can be a powerful tool to boost efficiency and knowledge, it is the human touch – those emotional and relational skills – that will make the real difference in the future of work.

REMUNERATION ON THE MANAGEMENT BOARD IN FINANCIAL INSTITUTIONS: CURRENT DEVELOPMENTS IN THE FRAMEWORK OF SUPERVISORY LAW, LABOR LAW, BEHAVIORAL ECONOMICS AND PRACTICE

JULIA REDENIUS-HÖVERMANN | Professor of Civil Law and Corporate Law and Director of the Corporate Governance Institute (CGI) and the Frankfurt Competence Centre for German and Global Regulation (FCCR), Frankfurt School of Finance and Management

LARS HINRICHS | Partner at Deloitte Legal Rechtsanwaltsgesellschaft mbH (Deloitte Legal) and Lecturer, Frankfurt School of Finance and Management¹

ABSTRACT

In the following article, selected topics in the current implementation of compensation systems for management boards are discussed in more detail, with the focus on the tension that regularly arises in compensation practice between the regulatory and labor law framework, behavioral economics, and (market) practice. To make the presentation more understandable, the regulatory legal bases generally refer to the requirements of CRD VI and cover topics that the national legislators of the individual E.U. member states have implemented in national law with the same content. It is shown that the practice of remuneration systems for management board members in institutions is based on a (mature) legal framework. Individual internal and external dynamic factors influence the further implementation of the remuneration systems for management board members and require a risk-based regular review process of the compatibility of the remuneration systems and their implementation with the regulatory requirements and the operational requirements of the institution, in particular from the updated business and risk strategy. Particularly, when it comes to the specific implementation of performance-related variable remuneration, institutions must take into account the dependence of regulatory requirements on the applicable labor and company law framework and reconcile these in a balanced and practicable manner. The question of whether the current (over)regulation will lead to a “regulatory infarction” in the near future remains to be discussed.

1. INTRODUCTION: THE TWO-TIER STATUS QUO FOR INSTITUTIONS OPERATING IN THE E.U.

From a legal perspective, the status quo on the content of management board remuneration in institutions in the E.U. currently presents a mixed picture:

1.1 The (mature) target picture of the sound compensation regulatory framework

Following several updates over the past 15 years, the (E.U.) legislator has set out a preliminary target for the core regulatory framework conditions for the content of remuneration, which

¹ We would like to thank Dorothea Langhans (research associate at Deloitte Legal) for her support in the preparation of this article.

the E.U. member states have each transposed into national law. The European Banking Authority (EBA) has fulfilled its mandate set by the E.U. legislator [Art. 74 (3) Directive 2013/36/EU, in the version of Directive 2024/1619/EU of May 31, 2024 (CRD VI)] to issue guidelines on sound remuneration policies with regard to the specification of the requirements set out in CRD VI on the content of remuneration systems and remuneration governance, most recently with the follow-up version of the guidelines on sound remuneration policies [EBA/GL/2021/04 of July 2, 2024, EBA-GSR 2.0] and the supervisory authorities of the individual E.U. member states have supplemented the EBA-GSR 2.0 for the application of the law with their own announcements for their own interpretation in their supervisory practice.²

The institutions domiciled in the individual E.U. member states (including the subordinate companies of institutions from other E.U./non-E.U. member states and branches of institutions domiciled in non-E.U. member states) have generally come to terms with the legal framework conditions – in particular with the two regulatory purpose considerations on which the legal framework conditions are generally based: (1) the monetary behavioral incentive of the individual employee of the institution and, thus, in particular, also of the management board member (= above all with regard to the incentive set by the variable remuneration components specifically granted), and (2) transparent risk management (= institution only grants affordable variable remuneration in line with the business and risk strategy and with full transparency for all relevant external and internal stakeholders).³ They have established the key regulatory requirements in the remuneration systems of the management board – and here above all in the variable remuneration components

– and have also created (mature) interactions between the relevant stakeholders in the implementation of the remuneration systems in the sense of needs-based remuneration governance.⁴

2.2 Establishment of needs-based (regular) communication between the institutions and the supervisory authorities as well as the auditors, including in the implementation of management remuneration systems – with feedback to the legislator

Supervision, auditing and remuneration practice have generally found a common starting point for regular communication. Among other things, the communication of rules entails an early exchange on the design of individual remuneration components and their implementation that require discussion. Coordination on the implementation of the remuneration systems (and in particular the possible granting of variable remuneration) in crisis situations at the institution is generally constructive, consensus- and solution-orientated.⁵ The auditors responsible for auditing the remuneration systems as part of the audit of the annual financial statements are usually involved by the institutions during the year in individual topics and problem areas where changes are required in order to reach a common understanding of a solution and implementation in the remuneration systems that complies with regulatory requirements. The legislator incorporates relevant experiences of remuneration practice from the implementation of the respective regulatory requirements into the remuneration systems and incorporates these into the subsequent amendments to the statutory requirements.⁶

² In Germany, BaFin's most recent announcement on the supervisory application of the Remuneration Ordinance for Institutions (Institutsvergütungsverordnung, IVV), which transposes the requirements of CRD VI into German law, "Questions and answers on the remuneration ordinance for Institutions (FAQ IVV)," of June 13, 2024, <https://tinyurl.com/y9tf4cet>

³ See Hinrichs, L., A. Kock, and D. Langhans, 2018, "Vergütung nach der Institutsvergütungsverordnung 3.0," <https://tinyurl.com/5f2hjctj>

⁴ This applies in particular to the division of labor between the remuneration control committee in its preparatory activities for the relevant resolutions and other decisions of the supervisory body in the implementation of the remuneration systems, which, above all, due to the increasing technical professionalization of the committee members, includes the technical discussion and debate of the relevant material topics required by the supervisory body prior to their resolution (e.g., in the assessment of the appropriateness of the remuneration systems, in their updating, e.g., in the variable remuneration parameters) to a sufficient extent.

⁵ This consensual approach is reflected above all in the supervisory side's intention to ensure a common understanding between the supervisory authority and the institution regarding the specific supervisory measures considered by the supervisory authority; for example, in cases where the supervisory authority imposes a cap or even prohibits and/or sets variable remuneration.

⁶ The privileged treatment of variable remuneration of risk takers with an annual amount of no more than €50,000 and no more than one-third of total remuneration, as stipulated by Directive 2019/878/EU (CRD V) in Art. 94 para. 3 CRD VI, resulted from the practical experience of remuneration practice that variable remuneration up to this quantitative amount does not require ex-ante risk adjustment from a standardized perspective through the partial granting over a retention period with malus and claw back testing and a partial granting in parameters aligned with the sustainable performance of the institution in accordance with Art. 94 (1) lit. l, m and o CRD VI in order to achieve the two purpose considerations.

This well-established status is flanked by the ongoing dynamic updating of the regulatory framework for individual content-related design parameters of the remuneration systems and their interaction with remuneration governance, which requires an ongoing and regular review with the verification of any need for modification in the content-related design of the remuneration systems and remuneration governance by the supervisory body and its implementation in remuneration practice; specifically, among other things, through:

- **Updates to the framework conditions directly related to the remuneration systems and remuneration governance:** even if remuneration is no longer the direct focus of legislative activities, the updating of individual legal framework conditions means that there is at least a need for readjustments to individual remuneration system design parameters. Currently, this results, among other things, from the legislative activities for the necessary updating of the remuneration strategy and its operationalization in the remuneration policy with regard to the institutions' risk appetite in relation to ESG risks in accordance with Art. 74 (1) lit. e) CRD VI with the necessary establishment of specific plans with quantifiable targets with regard to the financial risks arising from the short-, medium- and long-term ESG targets in accordance with Art. 76 (2) sentence 2 CRD VI.
- **Increasing complexity of the regulatory and market-related framework conditions for the proper business organization of institutions and their influence on remuneration:** the continuing increase in the complexity of the proper business organization of institutions requires, among other things, a constant further development of the risk strategy and, in particular, risk management and an associated constant increase in the fit and proper requirements for the professional suitability and reliability of managers. This has an impact both on the quantitative level of remuneration for the management board and on the specific structure of the individual remuneration parameters, particularly in the performance parameters of variable remuneration.

- **Focus of the supervisory body's activities on the appropriateness of the specific amount of the total remuneration of the management board members:** the constant – generally quantitatively increasing – development of the total remuneration for the individual management board members requires, among other things, that the assessment of the appropriateness of the specific amount of the total remuneration forms a continued focus of the activities of the supervisory body in the implementation and review of the remuneration systems. The sensitization of the respective supervisory body to ensure quantitative appropriateness results above all from the liability risks of the individual member of the supervisory body from the inappropriate (= unjustifiable in terms of the amount with regard to the relevant assessment parameters) total remuneration⁷ and necessitates an increasing concentration of the audit cycles for the regular appropriateness review. Recently, several institutions have begun to condense the regular audit cycle to a period of three years.
- **External influences on the market of institutions and their remuneration systems in the E.U.:** these external influences result from the relaxation of the remuneration law framework in the U.K. following the implementation of Brexit⁸ and also from the intensification of the hiring competition for suitable managers between the individual institutions in the institutional market and also between institutions and companies in the financial services sector that are not subject to any special regulatory requirements for the remuneration systems of managers (e.g., private equity market) or, from the perspective of the individual managers, are subject to more generous regulatory requirements compared to the institution-specific requirements.⁹
- **Active involvement of the shareholders/owners of the institution in the organisation of the remuneration systems:** recently, this active involvement in listed institutions has been driven in particular by proxy advisors for shareholders who, in annually published statements (policy guidelines), set out specific expectations

⁷ This applies, for example, to the supervisory board of institutions in the legal form of a stock corporation based in Germany under Section 93 AktG.

⁸ With effect from 31 December 2023, the legislator abolished the upper limit of 200% for the ratio between variable remuneration and fixed remuneration, which until then also applied to institutions domiciled in the U.K. (<https://tinyurl.com/4sxfujpn>).

⁹ For example, the regulatory requirements for the remuneration systems of the management board members of medium-sized investment institutions in accordance with the national implementing legislation of Directive 2019/2034/EU (IFD) and of capital management companies in accordance with the national implementing legislation of Directives 2011/61/EU (AIFD) and 2014/91/EU (UCITS), there is no absolute upper limit for the ratio between fixed and variable remuneration and, moreover, a more generous option than the requirements for banks to apply the principle of proportionality under supervisory law with the possibility of opting out of individual regulatory requirements on variable remuneration (e.g., on deferral, on variable remuneration). For example, on deferral, the application of malus and claw back regulations and the establishment of the remuneration component based on the sustainable performance of the institution (NWE component)].

and requirements for the content of the remuneration systems, particularly with regard to performance-based variable remuneration¹⁰ and thus significantly influence the voting behavior of shareholders with regard to the “say on pay” decision on remuneration in accordance with Art. 9 of Directive 2017/828/EU and the relevant national legal implementation regulations.^{11,12}

This article takes this current status quo as an opportunity to discuss individual selected topics in the current implementation of remuneration systems for management board members in more detail, with a focus on the tension that regularly arises in remuneration practice between the regulatory and labor law framework, behavioral economics and (market) practice. The regulatory legal bases generally refer to the provisions of CRD VI – for reasons of a more comprehensible presentation – and cover topics that the national legislators of the individual E.U. member states have transposed into domestic law with the same content.¹³

2. FIVE BASIC PARAMETERS FOR THE REMUNERATION SYSTEMS OF MANAGEMENT BOARD MEMBERS AND THEIR CURRENT ADAPTATION IN REMUNERATION PRACTICE

In practice, the implementation of remuneration systems and the remuneration governance of management board members continue to focus primarily on the following five basic parameters for the content of variable remuneration and can be summarized with the following practical implications:

2.1 Remuneration strategy as strategic implementation of the business and risk strategy in the remuneration systems – and its operationalization with the fixed and variable remuneration components and the respective specific remuneration parameters

The remuneration strategy forms the strategic core and starting point for the regulatory-compliant and supervisory-compliant implementation of the legal framework for the content of the remuneration systems and remuneration governance [Art. 74 (1) CRD VI]. It must ensure that the institution’s objectives

set out in the business strategy and the regulatory-compliant management of the risks arising from the implementation of the business strategy in accordance with the risk strategy are (also) implemented in the remuneration systems for the management board members – not only reactively by ensuring that the remuneration strategy is in line with the business and risk strategy of the institution, but ideally also actively from a regulatory perspective by setting remuneration parameters that promote the business and risk strategy.

Institutions must update their remuneration strategy to implement this regulatory requirement. To this end, the supervisory authorities require a standardized process that includes regular risk measurement and assessment of the impact of the management’s activities on the institution’s risk profile.¹⁴ This risk-sensitive management – as a component of risk management – must include the interaction between the management function of the management board members and the risk profile of the institution and the updating of the needs-based monetary behavior management through the remuneration system. It is operationalized by defining the individual fixed and variable remuneration parameters in the remuneration system of the management board and their respective specific remuneration parameters; in particular in the specific structure of the performance-based variable remuneration. Changes to the business and risk strategy can then require either the readjustment of existing remuneration components and/or remuneration parameters or the – temporary – introduction of new remuneration components and/or remuneration parameters.

Current practical examples of such situational temporary remuneration components include new functional allowances, which can be introduced as fixed remuneration components from a regulatory perspective if they remunerate a temporarily assumed more demanding task, function or organizational responsibility in addition to the regular function under the employment contract. From a regulatory perspective, the introduction of such a functional allowance for the management board is associated with the challenge that the overall responsibility of the institution’s management board

¹⁰ See the current versions of the policy guidelines of Glass Lewis (<https://tinyurl.com/3pu7smky>) and ISS (<https://tinyurl.com/ewhh9r87>), which are very present in remuneration practice.

¹¹ In Germany, for example, regulated in Section 120a AktG. For more details on Section 120a AktG, see BeckOGKAG/Hoffmann, edition as at 01.02.2022, Section 120a AktG, margin no. 1 et seq.

¹² On this and on individual listed companies where the shareholders have rejected the management board members’ remuneration system on the basis of a corresponding recommendation by the proxy advisors, see only Lünstroth, P., and T. Blumenstein, 2022, “Vorstandsvergütung auf verpflichtend auf dem Prüfstand,” <https://tinyurl.com/436h6t3w>

¹³ An overview of the domestic supervisory regulations adopted by the individual E.U. member states to implement the requirements of the CRD is published on the EBA website (<https://tinyurl.com/446vesev>)

¹⁴ See para. 199 EBA-GSR 2.0.

generally encompasses all operational and strategic topics, regardless of whether these are part of regular operations or have a temporary impact on the institution due to extraordinary internal or external factors – this makes it difficult in practice to distinguish between the “temporarily more demanding” task remunerated with the functional allowance and the fixed (basic) remuneration determined for the overall responsibility of the institution. It is, therefore, only permissible under supervisory law in individual cases if, for example, in the case of a departmental distribution of the individual areas of responsibility in the overall board of directors, individual management board members take on special additional operational tasks that result primarily from external influences and can be sufficiently clearly distinguished from the regular activities in the overall board-related responsibility. Recent practical examples include the granting of a functional allowance for institutions affected by Brexit, including for the extensive coordination with the respective supervisory authorities in the readjustment of business operations associated with Brexit.

This case-by-case adjustment of board remuneration is to be welcomed from a behavioral economics perspective, as it can restore the all-important “alignment of interests” in the context of the existing “principal-agent conflict”.¹⁵ The diverging interests of headmasters and agents can otherwise have a negative impact, particularly in the event of “information asymmetries”. Examples of how information asymmetries can have a negative impact arise in the situations in question, for example in the form of “hidden actions” or “hidden information”. This can be the case if board members act in a way that benefits them on the basis of information made available only to them but not to the headmasters,¹⁶ but use this advantage against the headmasters.

2.2 Determination and derivation of the performance-based remuneration parameters of the variable remuneration from the levels of institution/group, organizational unit and individual performance contributions of the individual management board member with a multi-year assessment period

The performance-based remuneration parameters of the variable remuneration operationalize the guiding principles of the remuneration strategy in the implementation of the remuneration system in the respective reference period. The regulatory requirements stipulate that the assessment parameters such as performance-related targets for the specific reference period must include the levels of the institution/group, the organizational unit, and the individual performance contributions of the individual manager (Art. 94 (1) lit. a) CRD VI) and that the performance assessment must generally take place within a multi-year framework that takes into account the business cycle and the business risks of the institution [Art. 94 (1) lit. b) CRD VI]. The individual targets must contain quantitative and qualitative performance criteria [Art. 94 (1) lit. a) CRD VI], whereby the EBA would also like to see the qualitative performance criteria applied to the levels of the institution and the organizational unit.¹⁷ These graduated regulatory requirements for the content of the performance-based remuneration parameters are intended to ensure that all relevant risks quantified in the on- and off-balance sheet items are taken into account in the measurement of variable remuneration across all financial years.¹⁸ The allocation of the relevant risks in the quantitative and qualitative targets at the individual levels is intended to ensure dedicated risk and behavior management.

In remuneration practice, the implementation of these regulatory requirements for the performance parameters of the management board members is associated with the challenge that the management board members, by virtue of their overall responsibility for the institution under company law, have a management-related responsibility for all significant risks. On the one hand, the organ-related overall responsibility comprises the factual level, according to which the management board must operationalize all substantive

¹⁵ Stadler, M., 2010, “Shareholder activism by hedge funds”; studies from the U.S. have shown that improvements in corporate governance help to reduce agency costs, Millstein, I. M., and P. W. MacAvoy, 1998, “The active board of directors and performance of the large publicly traded corporation,” *Columbia Law Review*, 1283, 1317 f.; see also: Siemens, P., 2023, *Die aktienrechtliche Entlastung*, Duncker & Humblot GmbH

¹⁶ In the stock corporation, the principals are only entitled to the information the agents provide to them, Redenius-Hövermann, J., 2019, *Verhalten im Unternehmensrecht*, Mohr Siebeck

¹⁷ Para. 231 EBA-GSR 2.0.

¹⁸ See para. 220 et seq. EBA-GSR 2.0

issues relating to the institution's business operations as part of operational management.¹⁹ On the other hand, it includes the committee-related level, according to which the individual management board members remain responsible for the overall operational management of the institution even if the individual tasks are allocated on a departmental basis – as is customary in practice – and are, therefore, subject to the organ-related duties to control and monitor the other management board members in the implementation of the management with regard to the departmental allocation of tasks.²⁰

This organ-related overall responsibility means that in individual cases, particularly at the organizational unit level, institutions are unable to allocate the relevant risks in a practical manner for behavioral and risk management as required by supervisory law and set suitable targets as assessment parameters for performance-based variable remuneration in order to achieve the aforementioned regulatory purpose. Against this background, individual supervisory authorities accept a combination of the organizational unit levels and the individual performance contributions for the determination of the assessment parameters, with the standardized view that the individual performance contribution and the performance contribution of the organizational unit (may) largely overlap.²¹ From a supervisory perspective, the combination of the levels of the institution's/group's objectives and the organizational unit is equivalent to such a cumulative determination of the individual target parameters with regard to the purpose of holistic risk allocation – it can even operationalize the overall responsibility of the body in an even more determined manner and control the mutual monitoring and supervision between the individual management board members of the institution and the individual departments.

From a behavioral economics perspective, care must be taken to avoid “short-termism” in this context. In particular, variable remuneration must not be based on targets that have too short an assessment period. In this respect, recommendation G.10 sentence 1 and G. 10 sentence 2 of the German Corporate Governance Codex (DCGK), which recommend predominantly share-based remuneration and a four-year holding period, are to be welcomed.²²

2.3 Ex-ante risk adjustment of variable remuneration: Delimitation of negative performance contributions from compensation, proper business judgment rule on the part of management in the assessment of negative performance contributions

According to the legislator's understanding, ex-ante risk adjustment – mirroring the determination of the remuneration parameters – involves considering the change in the relevant risks (up to their occurrence in individual cases) in the respective reference period when determining the performance-based variable remuneration. It essentially comprises, in accordance with Art. 94 (1) lit. n) CRD VI, for the respective reference period: (1) assessment of the specific target achievement for the individual assessment parameters (and here in particular a negative deviation from the agreed targets, “negative target achievement”), (2) assessment of the development of risks and their impact on the economic and financial performance of the institution (and here in particular on the earnings situation), and (3) actions of the individual management board member that are incompatible with the risk strategy and risk management of the institution, insofar as these have had an impact on the risk profile of the institution (negative individual performance contributions of the management board member). If such risks have materialized primarily in the form of negative performance contributions by the institution and/or negative individual performance contributions by the individual management board member, institutions should make a risk adjustment as a reduction in the (total) amount of variable remuneration, which in individual cases may lead to a complete cancellation of the variable remuneration in the respective reference period if the management board member was involved in, or responsible for, behavior that led to significant losses for the institution (Art. 94 (1) lit. n) sentence 5 lit. (i) CRD VI) or has not met the appropriate standards in terms of suitability and behavior [Art. 94 (1) lit. n) sentence 5 lit. (ii) CRD VI].

In remuneration practice, the implementation of these regulatory requirements from a labor law perspective is associated with the challenge, particularly for the assessment of any individual negative performance contributions of the individual management board member, of assessing

¹⁹ See Fleischer, H., m 2003, “Zum Grundsatz der Gesamtverantwortung im Aktienrecht,” *Neue Zeitschrift für Gesellschaftsrecht* 6, 449-459, on institutions domiciled in Germany in the legal form of a stock corporation with Section 77 (1) s. 1 AktG as the relevant legal source.

²⁰ See instead of all BeckOGK AktG/Fleischer, edition as at 02/2024, Section 77 AktG margin nos. 53 et seq.

²¹ See BaFin FAQ IVV, question 15.

²² Kremer, T., G. Bachmann, M. Lutter, A. von Werder, and H.-M. Ringleb, 2023, *Deutscher Corporate Governance Kodex*, 9th edition, Beck C. H., G.10 para. 2; see also Siemens (2023)

any reduction in variable remuneration to be made under supervisory law in conjunction with and, at the same time, in distinction to any claims for damages of the institution against the management board member resulting from the (co-) involved or responsible conduct of the management board member and, in the present context, essentially resulting from the breach of the management board member's organ-related duties to properly manage the business.²³ The interplay and delimitation must be assessed on the basis of the purpose of the ex-ante risk adjustment and the granting of corresponding organ-related claims for damages by the institution against the management board member: the ex-ante risk adjustment is intended to ensure the holistic consideration of the change in the allocated risks in the reference period on the variable remuneration and, according to the supervisory model of the legislator and the supervisory authority, should, therefore, be carried out at the starting point in all manifestations and here in particular for any negative individual performance contributions of the individual management board member regardless of fault.²⁴

In contrast, the award of the institution's organ-related claim for damages against the individual management board member due to an organ-related breach of duty is based on a compensatory function and is intended to protect the company's assets, and thus ultimately also the shareholders and creditors of the institution, from losses incurred by the institution due to a breach of the organ-related duties of care of the individual management board member in the management of the business.²⁵ It is subject to the principle of fault²⁶ in view of the management of the assets of the institution's shareholders as third-party assets associated with the organ-related management and the associated obligation to protect the interests of the individual stakeholders, which in addition to the shareholders also include the employees and the general public, and, therefore, requires at least a negligent or intentional breach of the duty(s) relevant to damages by the management board member (fault-based organ-related liability of the management board member). In this respect,

the compensation function of the organ-related claim for damages relating to the protection of the institution's assets overlaps with the ex-ante risk adjustment.

The relevant legal bases under company/labor law for fault-based liability of directors and officers provide the individual director with a liability-free entrepreneurial decision-making scope in accordance with the business/legal judgment rule, according to which the breach of duty giving rise to liability does not exist if the individual director makes the specific management decision on the basis of an uncertain factual situation (business judgment rule) or a legally ambiguous legal situation (legal judgment rule), an ambiguous legal question that is controversial in its legal application and has not been clarified by the highest court (legal judgment rule) on the basis of appropriate decisions for the benefit of the institution.²⁷ If the specific management decision within the framework of the business/legal judgment rule causes damage to the institution in the further course of time, the management board member can counter the proper application of the business/legal judgment rule with the institution's claim for damages.²⁸ In view of the overlap between the protective purposes of ex-ante risk adjustment and the compensatory function of the institution's organ-related claim for damages against the management board member, there are strong reasons from a teleological perspective to extend the scope of application of the business/legal judgment rule to individual negative contributions to success by the management board member that include the management board member's involvement in a fact/action, which has led to significant losses for the institution and, in this case group, to deny a reduction of the variable remuneration to the extent of the complete cancellation of the variable remuneration for the relevant reference period if the management board member can invoke the proper application of the legal/business Judgment rule with regard to the management decision relevant to the participation that resulted in the occurrence of the significant loss at the institution.

²³ The legal basis for such claims for damages for institutions based in Germany in the legal form of an AG/SE is Section 93 AktG (in conjunction with Art. 51 Regulation (EC)/2157(2001) (SE Regulation)), Art. 51 VO (EG)/2157(2001) (SE-VO) or in the legal form of a GmbH Section 43 GmbHG.

²⁴ See Buscher, A. M., C. von Harbou, V. Link, and T. Weigl, 2018, *Verordnung über die aufsichtsrechtlichen Anforderungen an Vergütungssysteme von Instituten*, 2nd edition, Schäffer-Poeschel, Section 18 InstitutsVergV marginal no. 119.

²⁵ See MüKoAktG/Spindler, 2023, 6th edition (<https://tinyurl.com/56eyf6td>), § 93 AktG marginal no. 1 on this protective purpose of the institution's claim for damages due to a breach of the duty of care of corporate bodies.

²⁶ See MüKoAktG/Spindler, Section 93 AktG marginal no. 5.

²⁷ On these legal principles for the application of the business/legal judgment rule instead of all MüKoAktG/Spindler, Section 93 AktG marginal no. 48 et seq.

²⁸ For this legal consequence of the proper application of the business/legal judgment rule, see MünchKommAktG/Spindler, Section 93 AktG marginal no. 46.

With regard to directors' and officers' liability, the behavioral control of the board member before any damage occurs must be emphasized from a behavioral science perspective. The assertion of claims for damages is crucial here. *De lege lata*, however, there are considerable enforcement deficits here.²⁹ In this respect, the use of claw backs to close this gap, which is not required by law but recommended in recommendation G.11 sentence 2 DCGK, is worthy of discussion. Furthermore, the reform of directors' and officers' liability should be considered.³⁰

2.4 Ex-post risk adjustment through malus and claw back: Labor law transparency of the possible case groups of claw back cases

The ex-post risk adjustment pursuant to Art. 94 (1) lit. n) CRD IV extends the consideration of the change in the relevant risks in accordance with the ex-ante risk adjustment (up to their occurrence in the individual case) in relation to the respective reference period of the variable remuneration over time (1) when determining the individual retained remuneration components of the variable remuneration granted for the respective reference period (malus test), and (2) when assessing whether the significant changes in the relevant risks over time also affect the variable remuneration components already paid out by the institution for the respective reference period and whether the management board member must repay all or part of the variable remuneration components already paid out (claw back test). The period-related assessment of the ex-post risk adjustment must (only) take into account the changes in risks identified at the time of the respective audit that relate to the specific reference period. The malus or claw back test of the relevant reference period, therefore, does not consider the identified change in relevant risks that relate to other reference periods. The ex-post risk adjustment, therefore, generally (only) covers cases in which the event giving rise to the risk occurs in the relevant reference period and the profit contribution resulting from the change in risk (in particular relevant negative profit contributions of the institution) occurs in the respective subsequent period before the malus or claw back test is carried out.

In remuneration practice, the implementation of ex-post risk adjustment (and in particular the claw back test) is associated with the labor law challenge for many institutions that the labor law framework in individual E.U. member states for contract design with the formal transparency requirement sets restrictive requirements for the content of claw back provisions in the employment contract and, in this context, requires in particular a specification of the relevant case groups in the employment contract that go beyond the abstract requirements of Art. 94 (1) lit. n) CRD VI. In addition, several jurisdictions within the scope of CRD VI have implemented the labor law principle that a remuneration component linked to the performance of work – which the performance-based variable remuneration for management board members already has in view of the mandatory individual targets to be established as performance parameters – is generally earned with the performance of the work and can no longer be withdrawn from the management board member.³¹ Against this background, any claw back claims are generally only asserted with restraint in remuneration practice in the relevant jurisdictions.

It must be ensured that any claw back or retention options, for example in the form of a reduction in management board remuneration in accordance with Section 87 (2) AktG, are also considered if the conditions of the offence are met.³² This is obvious in the context of liability claims due to the interest in restitution, but must also be emphasized at the same time due to its effect on behavior.

2.5 Maximum upper limit for variable remuneration

The regulatory requirements stipulate a general upper limit of 100% for the ratio between the fixed remuneration components and the variable remuneration components [Art. 94 (1) lit. g) (i) CRD VI]. This upper limit can be increased by the shareholders of the institution by resolution to a maximum of 200%, whereby the resolution must be based on a recommendation by the institution (= the management board and the supervisory body) with comprehensive documentation of the reasons and the expected impact of the higher upper

²⁹ On the problem of the lack of enforcement of liability claims Redenius-Hövermann/Siemens, ZIP 2020, 145 et seq.

³⁰ Redenius-Hövermann, J., 2024, "Der Aufsichtsrat," pp. 84, 86, which discusses the initial impetus for the reform of directors' and officers' liability.

³¹ See for German labor law judgment of the German Federal Labor Court dated November 13, 2013, 10 AZR 848/12.

³² On a deficit in this regard, which has become particularly apparent in the context of the COVID-19 pandemic: Redenius-Hövermann, J., and P. Siemens, 2022, "Vorstandsvergütung und ESG – Auswirkungen von ARUG II, Corporate Finance Sonderheft "ESG und Konsequenzen für Unternehmensfinanzierung und Finanzanlagen", 05-06/2022, S. 140 ff. ZIP 2020, 1585 et seq. In this context, *de lege ferenda*, the mandatory application of the regulation could also be appropriate.

limit on the requirements for a sound capitalization of the institution [Art. 94 (1) lit. g) (ii) CRD VI]. These substantive requirements for the draft resolution are intended to ensure that the institution's basic financial resources are in line with regulatory requirements (especially with regard to capital adequacy) even if such quantitatively higher variable remuneration is granted to the managers (risk management function) and that the shareholders can make the decision to increase the cap in full knowledge of, among other things, the quantitative risk-related effects on the institution's capital adequacy and the associated risk-bearing capacity (transparency function).³³ In order to ensure the continued fulfillment of these two functions, the continued validity of the resolution and the recitals documented in the draft resolution must be regularly reviewed and a new resolution must be passed if necessary.

In practice, it is primarily listed institutions and other institutions organized under private law with cross-border activities in capital market-oriented business areas that apply a higher cap in the E.U. and then generally use the maximum permissible limit of 200%. One of the challenges for these institutions is keeping track of the specific reasons for the increased upper limit, particularly if the number of beneficiary employees changes over time. Suitable reasons for the increased upper limit can be derived from the business strategy and the associated remuneration strategy as well as from a competitive perspective. The remuneration strategy may provide for an increased upper limit in particular if the variable remuneration system provides for a multi-year participation program with real or virtual shares in the institution in addition to the general performance-based remuneration; this is particularly the case in growth business strategies in which the institution wishes to incentivize the individual manager with a benefit from the participation program and/or wishes to retain the manager in the long term by structuring the program in line with requirements (retention element).

The retention element is also the focus when deriving the specific reason from a competitive perspective. Over time, the institution must demonstrate that the specific reasons underlying the decision to increase the cap also exist during the implementation of the remuneration system; with regard to the transparency function, a regular process is required from a regulatory perspective that includes the regular review of the specific reasons, the number of beneficiary employees, and the impact of the higher (cash) benefits resulting from the higher cap on capital adequacy. If the specific reasons no longer exist, the institution must examine whether there are other reasons suitable for a higher cap in relation to the beneficiary managers and adopt a new resolution to increase the cap by the shareholders. A new resolution is also required if the – operational – parameters for capital adequacy change and, therefore, in particular in the event of a reduction in capital adequacy over time, the higher remuneration payments resulting from the increased cap – still – do not have a negative impact on them from a regulatory perspective. In addition, when implementing the increased cap, institutions must generally bear in mind that the higher total remuneration resulting from the higher variable remuneration component gives rise to corresponding expectations among managers, particularly when negotiating the follow-up employment contract for the subsequent appointment period, and that an increase in the fixed remuneration components of the total remuneration may, therefore, become relevant if the cap is reduced as required from a regulatory perspective.

From a behavioral science perspective, it is important to ensure that the target shifting effect, or the self-service effect and group behavior are contained. Control by the annual general meeting or the supervisory authority as well as transparency through the disclosure of remuneration can contribute to this. However, even behavioral research will not be able to determine an exact maximum limit up to which remuneration can still have a performance-enhancing effect. It must also be ensured that other behaviors, such as the lift effect, are not reinforced by disclosure.³⁴

³³ See para. 48 EBA-GSR 2.0.

³⁴ Redenius-Hövermann, J., 2019, "Verhalten im Unternehmensrecht," p. 106 et seq. with further references.

3. CONCLUSION

The practice of remuneration systems for management board members in institutions is based on a (fully) developed legal framework and generally on a common basic understanding of the institution with the supervisory authority and the auditor regarding the specific content requirements, which generally provides the individual institution with reliable planning security in the further implementation of the remuneration systems from a regulatory perspective.

At the same time, individual internal and external dynamic factors influence the further implementation of the remuneration systems for the management board members and continue to require a risk-compliant regular review process for the compatibility of the remuneration systems and their implementation with the regulatory requirements and the operational requirements of the institution, in particular from the updated business and risk strategy.

In particular, in the specific implementation of performance-based variable remuneration, institutions must observe the dependency of the regulatory provisions with the applicable labor law and corporate law framework parameters and bring these into a balanced and practical harmony. To this end, the individual institution must reflect on the relevant legal considerations for the specific implementation of the regulatory requirements within the labor and company law framework and document them in an appropriate manner (above all in relevant legal opinions).

In terms of legal policy, it remains to be discussed whether the current (over)regulation will lead to a “regulatory infarction”³⁵ in the near future and whether national and European regulators are, therefore, urgently called upon to make adjustments by way of deregulation.

³⁵ Roland Koch recently coined this term, see 75 Jahre Grundgesetz – 75 Jahre Soziale Marktwirtschaft – LUDWIG-ERHARD-STIFTUNG E.V. (<https://tinyurl.com/52hsxusn>) or Mit Planwirtschaft wird Klimapolitik scheitern – LUDWIG-ERHARD-STIFTUNG E.V. (<https://tinyurl.com/4ym4u3e3>). Wolfgang Schön uses the term of “Regulierungsbankrott”, see Fachkräftemangel und Überforderung steigen: Bürokratie in Deutschland (faz.net) (<https://tinyurl.com/b5sm6jtd>).

© 2024 The Capital Markets Company (UK) Limited. All rights reserved.

This document was produced for information purposes only and is for the exclusive use of the recipient.

This publication has been prepared for general guidance purposes, and is indicative and subject to change. It does not constitute professional advice. You should not act upon the information contained in this publication without obtaining specific professional advice. No representation or warranty (whether express or implied) is given as to the accuracy or completeness of the information contained in this publication and The Capital Markets Company BVBA and its affiliated companies globally (collectively "Capco") does not, to the extent permissible by law, assume any liability or duty of care for any consequences of the acts or omissions of those relying on information contained in this publication, or for any decision taken based upon it.

ABOUT CAPCO

Capco, a Wipro company, is a global management and technology consultancy specializing in driving transformation in the energy and financial services industries. Capco operates at the intersection of business and technology by combining innovative thinking with unrivalled industry knowledge to fast-track digital initiatives for banking and payments, capital markets, wealth and asset management, insurance, and the energy sector. Capco's cutting-edge ingenuity is brought to life through its award-winning Be Yourself At Work culture and diverse talent.

To learn more, visit www.capco.com or follow us on LinkedIn, Instagram, Facebook, and YouTube.

WORLDWIDE OFFICES

APAC

Bengaluru – Electronic City
Bengaluru – Sarjapur Road
Bangkok
Chennai
Gurugram
Hong Kong
Hyderabad
Kuala Lumpur
Mumbai
Pune
Singapore

MIDDLE EAST

Dubai

EUROPE

Berlin
Bratislava
Brussels
Dusseldorf
Edinburgh
Frankfurt
Geneva
Glasgow
London
Milan
Paris
Vienna
Warsaw
Zurich

NORTH AMERICA

Charlotte
Chicago
Dallas
Houston
New York
Orlando
Toronto

SOUTH AMERICA

São Paulo

THIS UNIQUE IMAGE WAS GENERATED USING MID-JOURNEY, STABLE DIFFUSION AND ADOBE FIREFLY

WWW.CAPCO.COM



CAPCO
a wipro company