

CAPCO

a **wipro** company

THE CAPCO INSTITUTE
JOURNAL
OF FINANCIAL TRANSFORMATION

ORGANIZATION

AI safety and the value
preservation imperative

SEAN LYONS



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 Birscher, 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 Gomer, 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 Langenbucher, 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.

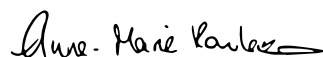
The authors in this edition offer fresh perspectives on the mindful use of GenAI and the implications of advanced GenAI on financial markets, in addition to providing practical and safe frameworks for boards and firms on how to approach GenAI governance.

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.

A handwritten signature in black ink, reading "Annie Rowland". The signature is fluid and cursive, with a long horizontal stroke at the end.

Annie Rowland, **Capco CEO**

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

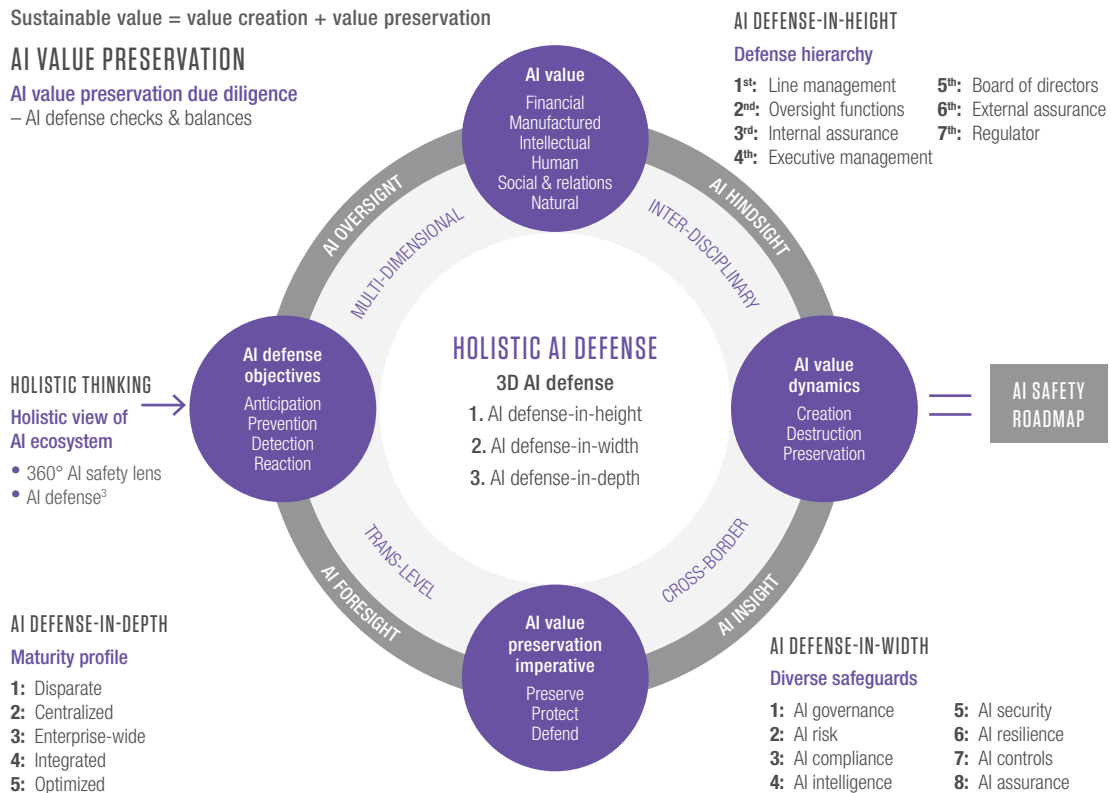
war. This potential for misuse and possible unintended catastrophic consequences could ultimately pose a threat to international security, global safety, and, ultimately, humanity itself.

- **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.

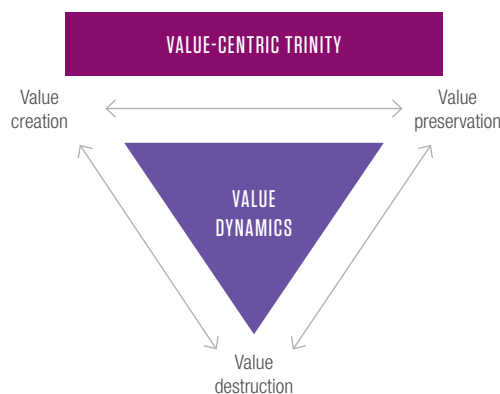
5.2 AI value dynamics

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)].

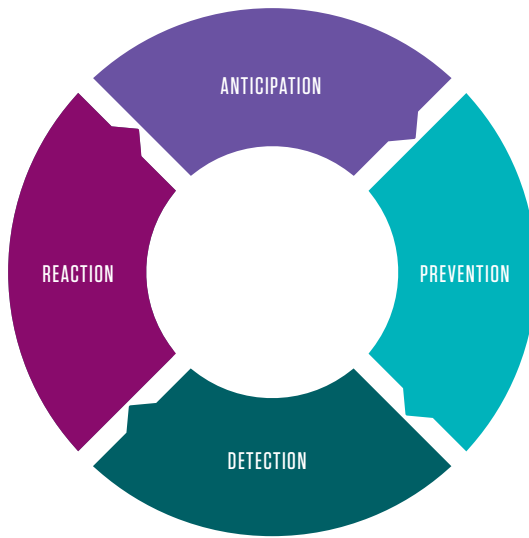
- AI value creation:** value creation is typically associated with enhancing value, increasing value, and generating value. Examples of how AI can create value for its stakeholders include efficiency and productivity, enhanced decision making, personalization, cost reduction, innovation, risk management, 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.

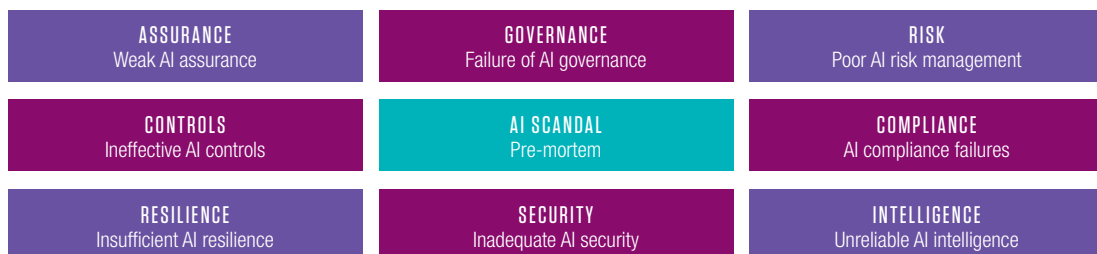
AI safety oversight by external gatekeepers and watchdogs can help to address the separation of power issue that is an inherent flaw present in self-regulation and in voluntary adherence. This can include various sources of external assurance (e.g., validation, certification, ratings, etc.) and

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/3jpxjeed>
- 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/y773yvj6j>
- CSA, 2024, "AI resilience: a revolutionary benchmarking model for AI safety," Cloud Security Alliance, May, <https://tinyurl.com/y7ijxyxy>
- 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/5u6ewrsn>
- 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>

© 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