INESC TEC

science-society.inesctec.pt

SCIENCE &SOCIETY

EDITION #07
JAN 2025

The responsibilities of Responsible AI

ABOUT US

INESC TEC SCIENCE & SOCIETY
PUBLISHED BY INESC TEC – INSTITUTE FOR SYSTEMS AND COMPUTER ENGINEERING, TECHNOLOGY AND SCIENCE
Campus da FEUP
Rua Dr. Roberto Frias
4200-465 Porto
Portugal
+351 222094000
info@inesctec.pt
www.inesctec.pt

ONLINE EDITION
science-society.inesctec.pt

INDEX

EDITION #07

This is the seventh issue of “INESC TEC Science & Society” magazine, a publication that seeks to communicate impactful science to a wide audience, this time addressing a topical question - the future of leadership in research and development in the era of Artificial Intelligence. This issue is different from the previous ones, focusing on disseminating the results obtained in the discussion held at an event organised by the “INESC Brussels Hub” in liaison with the European Commission and the NCBR Office, in Brussels. In June 2024, 60 people - including managers, researchers and policymakers - gathered in Brussels to discuss the future of leadership in R&D in the era of AI.

Artificial Intelligence has been changing the paradigm of leadership in R&D. This is precisely what we seek to show in this issue, through the opinions of external experts and INESC TEC researchers. We begin this issue by providing a broader view of the topic, then moving on to the discussion of sectoral applications. I will leave the details for the following editorial.

In addition to the magazine, we’ve decided to explore other media; over the last year, we have revisited – in chronological order of publication – older issues, exploring them through podcasts and videocasts.

We hope to keep fulfilling our goal of broadening the dissemination of science to a wider audience.

ARTUR PIMENTA ALVES
Series Coordinator
Emeritus Professor at FEUP
INESC TEC Director
artur.p.alves@inesctec.pt

"What is the future of R&D leadership in the era of Artificial Intelligence?"
an event organised by the "INESC Brussels Hub" in liaison with the European Commission and the NCBR Office, in Brussels, in June 2024

EDITORIAL

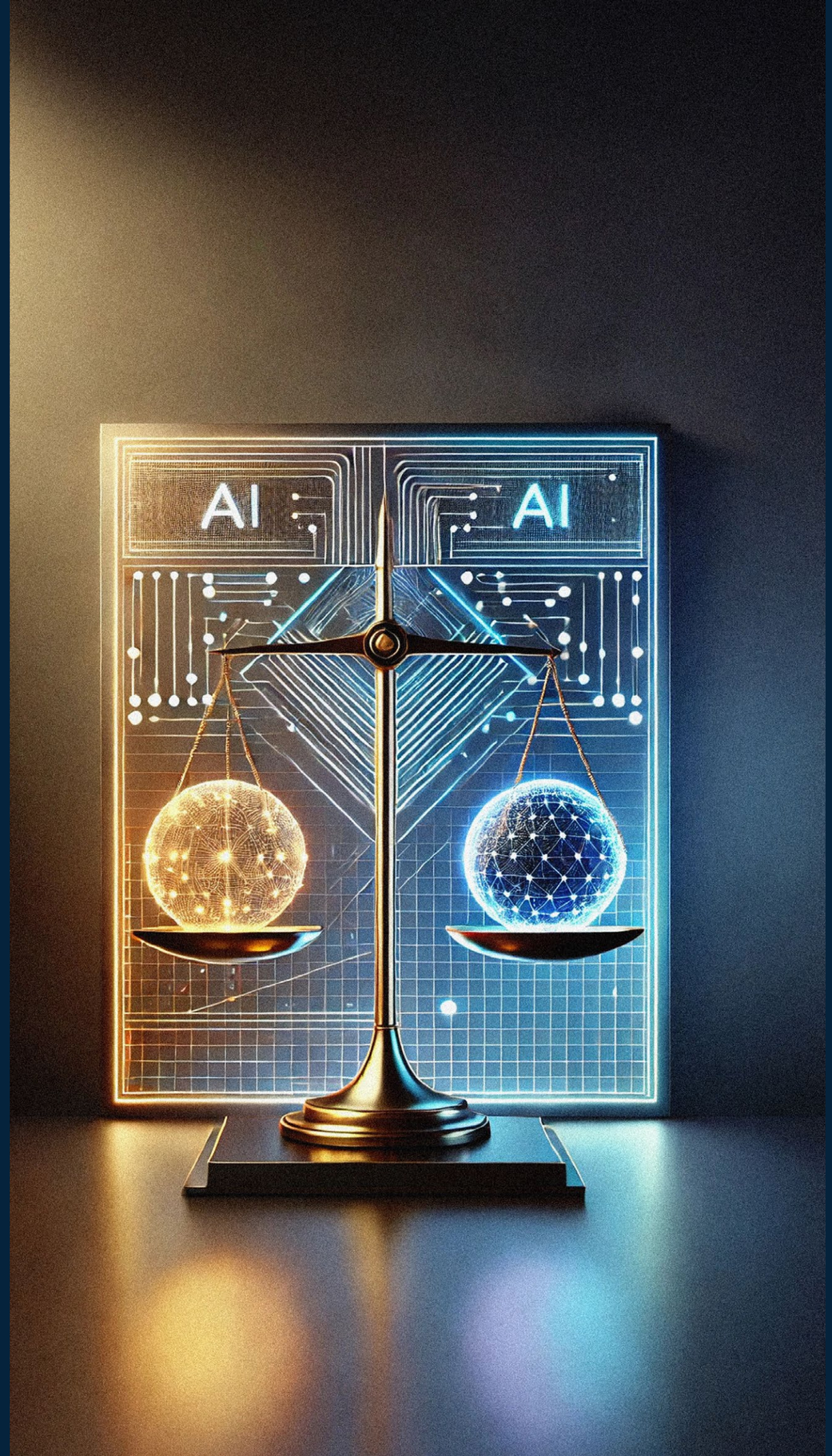


Image generated with AI using OpenAI's DALL-E

Artificial Intelligence (AI) is a constantly evolving scientific domain. It's hard to predict what the future holds for this kind of technology. However, it's clear that AI is changing our lives in significant ways. It's shifting the way we work, live, and make decisions. It is currently going through a phase of fast growth and development.

LIA PATRÍCIO (1, 2)
lia.patricio@inesctec.pt

JOÃO GAMA (1)
joão.gama@inesctec.pt

(1) Institute for Systems and Computer Engineering, Technology and Science (INESC TEC)
(2) Faculty of Engineering of University of Porto (FEUP)

A future with responsible AI must be collaborative. It will require a close cooperation between states, companies, and society. In the upcoming years, AI is set to have an ever-increasing impact on our lives, in areas like healthcare, education, work, and mobility.

Looking back, today's present issues have been addressed by remarkable technological advances. Currently, we face issues with major social impact: climate change and global warming, population growth and food production, desertification and water control. AI is contributing to and will play a key role in developing sustainable solutions to these problems. The changes associated with AI are likely to be far more impactful than any other technological revolution in human history.

Depending on the direction this revolution takes, AI will strengthen our ability to make more informed choices or reduce human autonomy; it will create new forms of human activity or make certain jobs redundant; it will help ensuring well-being among the many or increase the concentration of power and wealth in the hands of a few; it will expand democracy in our societies or endanger it. Understanding and addressing these tensions is of particular importance at a time when Europe perceives AI as a window of opportunity to overcome innovation and productivity gaps, while preserving the objectives of equity and cohesion as outlined in the Draghi report.

The choices we face today relate to fundamental ethical questions about the impact of AI on society; in particular, how it affects work, social interactions, healthcare, privacy, justice, and safety. The ability to make the right choices requires new solutions to fundamental scientific questions in AI and human-machine interaction. Choices that must be made today.

This issue of the magazine provides multiple outlooks on the opportunities and challenges associated with AI, and how to support more informed choices.

The growing impact of AI on society reinforces the importance of responsible AI, as all authors pointed out. Nuno Paiva states that "technology is not ethically neutral; it significantly shapes our values, behaviours, and societal norms." However, opinions differ on the ways and processes of obtaining a responsible AI.

Virginia Dignum mentions that "current AI systems, particularly Generative AI (GenAI), are designed to prioritize immediate performance over long-term maintainability and ethical considerations." Adopting an ethical relational approach to AI, combined with a structured, engineering-focused compositional paradigm, will lead to the development of AI systems that are not only powerful and efficient but also aligned with human values and societal needs."

Pedro Saleiro states that "there’s no shortage of research on fairness, safety, robustness, explainability, and privacy in AI, but there’s still a big gap between research work and real-world practice." Pedro stresses the need to regulate, test and evaluate not only the results of AI tools but also the processes that lead to these results. "Testing is the only way to ensure that AI behaves reliably in a range of environments, including edge cases that could present significant risks. A good analogy is to look at mission-critical industries like aerospace or nuclear energy, where failure is not an option. In these sectors, thorough testing is built into every stage of the development process, from initial design to final implementation. AI should be no different." The recent initiatives of the European Commission to regulate AI (namely the AI Act), despite being globally positive, present certain risks in terms of effective implementation - especially if they focus "more on legalistic compliance - producing mountains of paperwork rather than ensuring that AI systems are thoroughly and comprehensively tested".

Nuno Paiva emphasises that responsible AI should be based on ethical considerations in line with society's values and legal norms. He proposes a framework for "Trustworthiness and Human-Centred Design" from existing research. According to Nuno, "trust is built when users perceive the system as fair, transparent, and reliable." These ethical considerations should not be perceived as a burden, or as a mere need for ‘compliance’. On a positive note, Nuno Paiva points to a positive relationship between performance and adoption of responsible AI practices by leading AI companies.

Several articles in this issue address the transformative opportunities of AI. Diana Viegas and Nuno Cruz illustrate the multiple ways AI can expand capabilities to explore the deep sea, including real-time monitoring and autonomous operation, so robotic systems can stay underwater for longer periods, and reach deeper into ocean exploration. In terms of power and energy, Ricardo Bessa highlights the potential of AI for the energy transition, by supporting decision-making in energy systems with a high renewable component, where flexibility is key; or even for optimising the operation of energy communities or electric vehicle charging.

Moreover, harnessing the potential of AI requires tailored strategies. Pedro Amorim and Gonçalo Figueira propose five pillars for organisations to effectively transform AI's potential into reality. These pillars involve combining the tasks with the appropriate AI tools; exploring different AI approaches; using explainable AI methods; enabling different modes of interaction with humans; and ensuring AI literacy in the organisations.

Based on his experience of using ChatGPT in teaching, José Nuno Oliveira highlights the need to develop critical thinking in the use of Large Language Models (LLMs), and the importance of these skills in the jobs of the future. Finally, António Batista and António Lucas Soares draw attention to the need to consider not only the benefits, but also the negative effects of AI, namely on global energy consumption and its climate effects, as well as the impact of AI on the workforce and on social pressures.

Again, these tensions require reflection and action, and an ability to make the right choices considering different perspectives. As mentioned by João Claro and Arlindo Oliveira, the opportunities and challenges of AI require leadership that balances innovation and responsibility, going beyond efficiency and competitiveness, and building a vision focused on societal challenges, and based on ethical and collaborative principles. This leadership is fundamental to make the necessary choices to leverage an AI based on European values, so it becomes a key driver of innovation and competitiveness.

AI AND LEADERSHIP – THE NEXT PHASE. ARE WE PREPARED?

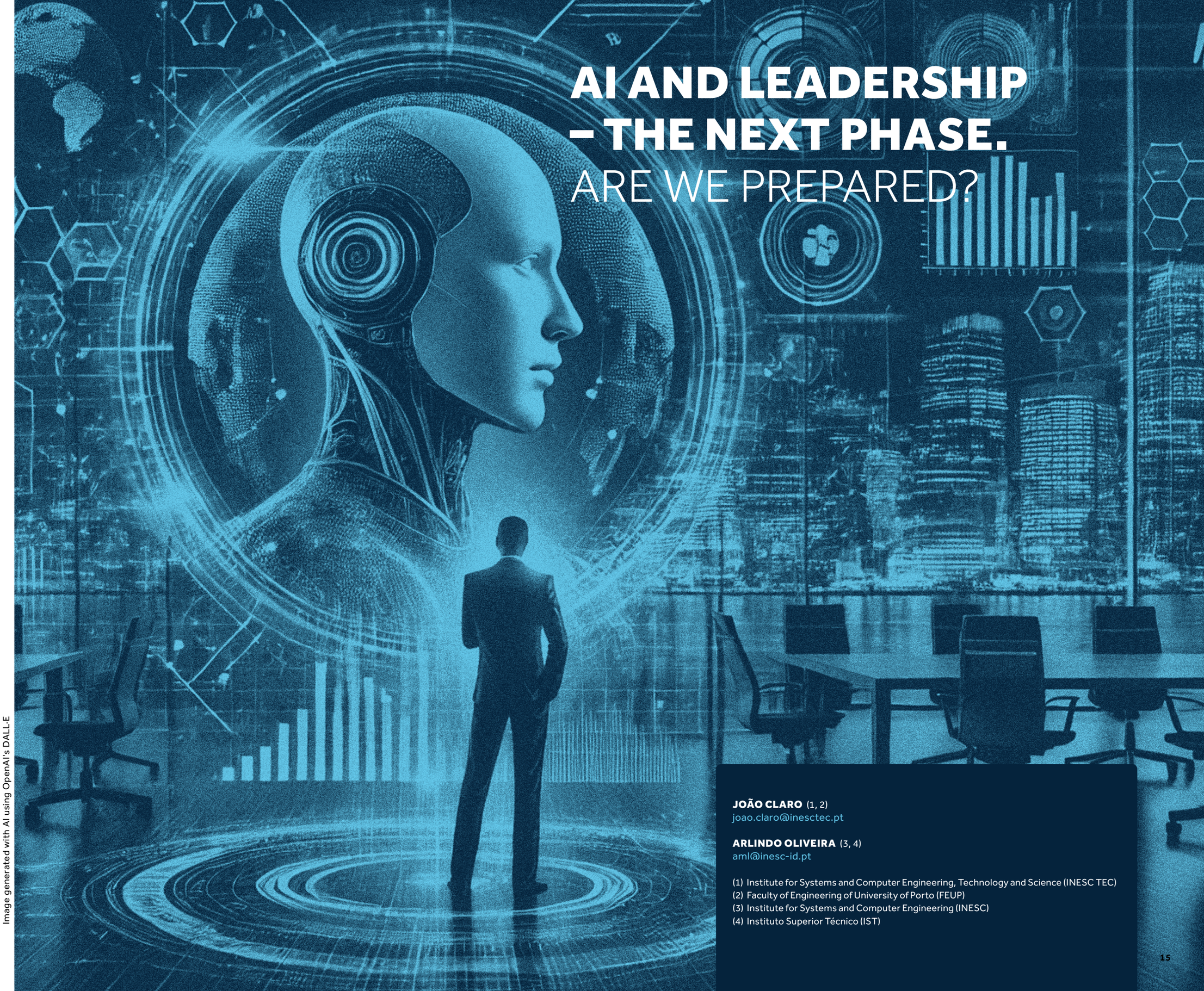


Image generated with AI using OpenAI's DALL-E

JOÃO CLARO (1, 2)
joao.claro@inesctec.pt

ARLINDO OLIVEIRA (3, 4)
aml@inesc-id.pt

(1) Institute for Systems and Computer Engineering, Technology and Science (INESC TEC)
(2) Faculty of Engineering of University of Porto (FEUP)
(3) Institute for Systems and Computer Engineering (INESC)
(4) Instituto Superior Técnico (IST)

As Artificial Intelligence (AI) continues to evolve, it is reshaping nearly every aspect of society – from the economy and healthcare to industry, education, science, and government. Research, development, and innovation (RDI) organisations, positioned at the forefront of this transformation, urgently need skilled leadership that can guide the adoption of AI while addressing ethical, collaborative, and broader societal implications. Effective AI leadership extends beyond enhancing organisational efficiency or competitiveness; it calls for a visionary approach that champions societal change, upholds ethical standards, promotes cross-disciplinary collaboration, and manages complexity in an unpredictable landscape. Leaders must balance innovation with responsibility, leveraging AI's potential to advance both organisational objectives and societal well-being.

This article explores the critical roles that RDI leaders must embrace to navigate these demands, drive transformation in their organisations, and create a lasting, positive impact.

ENABLING SOCIETAL CHANGE BEYOND ORGANISATIONAL BOUNDARIES

To fully harness the potential of AI, leaders of RDI organisations must extend their vision beyond institutional objectives to consider the ethical and societal implications of their work. AI presents unparalleled opportunities to tackle complex global challenges, from climate change to public health. However, achieving these outcomes responsibly requires leaders to foster cross-organisational collaboration, establish ethical standards, and drive societal change that prioritises public welfare.

For RDI leaders, ethical stewardship should be a core element of their mission, guiding practices that not only advance research and innovation, but also reflect values of fairness, transparency, and accountability. Rather than focusing solely on proprietary advancements, leaders need to engage with governments, academia, and industry stakeholders to shape responsible AI policies and favour projects that benefit the public good.

A significant challenge is that many in academia and RDI organisations are not fully prepared for this shift. Unlike traditional research, which may take years to influence society, AI demands rapid adaptation and agile decision-making. Leaders must embrace open, collaborative ecosystems where knowledge and resources are shared freely. This shift requires rethinking competitive boundaries and exploring new partnership models that align organisational goals with broader societal needs.

EXPANDING THE INSTITUTION'S ROLE EXTERNALLY AND INTERNALLY

For RDI organisations to lead effectively in AI, leaders must broaden their perspective to address both the organisations' external influence and internal culture. Externally, leaders should advocate for responsible and ethical AI practices, engage in policy discussions, and contribute to public education on AI's potential and risks. By establishing themselves as thought leaders in AI ethics, privacy, and data protection, these organisations can set standards that benefit society and reinforce their credibility as trusted pioneers in advancing AI for the public good.

Internally, fostering an environment of openness and experimentation is essential. Leaders must support a workplace culture where people feel empowered to share ideas, take risks, and explore new approaches without fear of failure. Given the uncertainties and rapid pace of AI development, a culture of experimentation is particularly valuable. Leaders should create collaborative teams where AI complements, rather than replaces, human expertise, encouraging researchers to blend creativity and critical thinking with AI's analytical strengths.

Advanced training for leaders and staff on technical aspects, ethics, and societal impacts of AI can help prepare organisations to manage AI responsibly. Access to external experts, interdisciplinary knowledge, and diverse viewpoints should be encouraged, enabling leaders to draw from a broad base of insights in decision-making. Leaders should also integrate diverse expertise – from data science to ethics – to enrich AI initiatives, ensuring that advances are informed, fair, and reflective of varied outlooks.

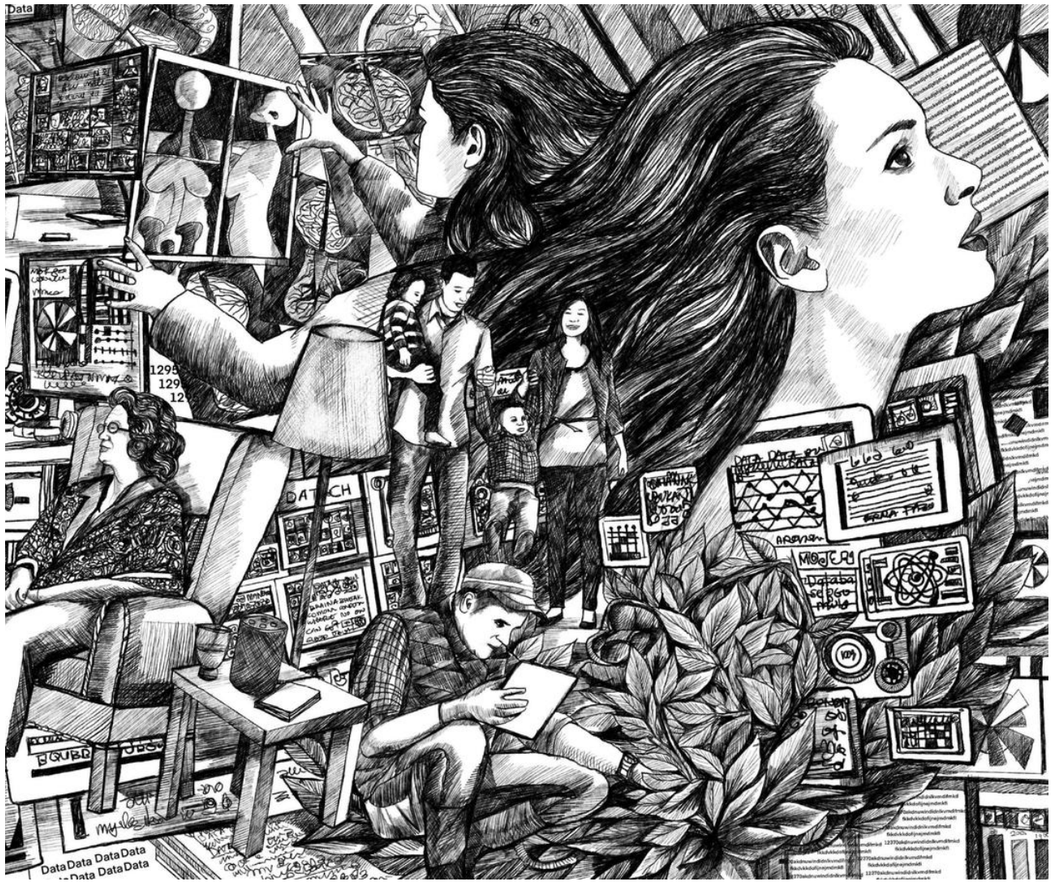
Furthermore, addressing non-technical concerns like ethics and privacy within the organisation requires a proactive approach. As AI evolves, so do ethical and privacy challenges. Leaders must work to embed ethical principles into the development and deployment of AI systems, ensuring that technologies are designed and implemented with fairness, transparency, and accountability in mind. This commitment to ethical standards not only minimises risk but also builds trust within the organisation and with the public.

MANAGING CHANGE IN AN UNCERTAIN ENVIRONMENT

In the rapidly evolving field of AI, uncertainty is a constant. Leaders in RDI organisations must excel at managing change in an environment where technological advancements and their implications are often unpredictable. A critical aspect of this is understanding and intensifying transition areas, from determining where to invest in AI technologies to anticipating shifts in the skill sets required. Leaders must balance short-term projects with long-term strategies, ensuring that resources are allocated effectively despite the risk of hardware and software obsolescence. By fostering collaboration between researchers and AI systems, leaders can create resilient teams where AI's strengths in data processing complement human insights and intuition.

The accelerated pace of technological obsolescence means that leaders must stay agile and continuously update their infrastructure. Investing in flexible, scalable technology solutions that can adapt to change is essential, as is cultivating a culture that embraces ongoing learning and adaptability. Research management practices also need to evolve to capture the unique dynamics of AI-driven innovation, moving beyond traditional metrics to emphasise flexibility, responsiveness, and the ability to pivot as new opportunities arise.

To address these challenges, leaders must foster an open culture that encourages experimentation and innovation. A safe environment for trial and error is vital, especially in a field where breakthroughs often stem from unconventional thinking. Leaders should encourage teams to approach problems creatively and support diverse perspectives to inspire new solutions. This commitment to continuous learning and flexible strategies is essential to equip researchers to manage the risks associated with AI, including preparing for technology obsolescence through scenario-based approaches to project management.



Ariyana Ahmad & The Bigger Picture / Better Images of AI / AI is Everywhere / CC-BY 4.0

Image generated with AI using OpenAI's DALL-E

PREPARING LEADERS FOR THE FUTURE OF AI

Given the academic community's current unpreparedness for AI's rapid transition, developing advanced training programmes for leaders is essential. These initiatives should equip leaders not only with technical expertise, but also with the ethical, collaborative, and strategic insights needed to navigate AI's organisational and societal challenges. Preparing leaders to foster effective human-AI collaboration is crucial; training should privilege skills like interdisciplinary team building, ethical foresight, and agile decision-making, enabling leaders to integrate AI in ways that support and enhance human expertise.

Leaders must also be skilled in assessing and mitigating the non-technical risks associated with AI, including ethical dilemmas, societal impacts, and potential biases. Regular assessments of AI projects from ethical, legal, and social perspectives can help organisations identify and address potential issues before they escalate. Additionally, training should prepare leaders to communicate AI's implications effectively to non-experts, fostering transparency and building public trust.

To stay informed, leaders should seek access to outside experts, participate in workshops on emerging technologies, and pursue opportunities for cross-sector learning. Encouraging a culture of continuous learning within RDI organisations is equally important, ensuring that leaders and their teams remain agile and informed as the AI landscape evolves. This commitment to ongoing development will prepare the next generation of AI leaders to address the complex, dynamic challenges of integrating AI into research and innovation.

NAVIGATING GEOPOLITICAL CHALLENGES IN THE GLOBAL AI LANDSCAPE

The global race in AI development presents unique geopolitical and ethical challenges that RDI leaders, especially those in Europe, must address to remain competitive while upholding high ethical standards. Rapid advancements in AI technology by the United States and China have positioned these countries as dominant players, setting high benchmarks in AI research, development, and commercialisation. This competitive landscape creates significant pressure for European RDI organisations to keep pace, not only in technological capabilities but also in shaping AI's ethical, social, and political frameworks. European leaders face a dual challenge: advancing AI technology within their organisations while upholding Europe's values of privacy, transparency, and inclusivity.

Positioning Europe as a leader in ethical AI provides an opportunity to set an international standard for responsible AI practices, advocating for global norms in privacy, transparency, and human-centred AI. European leaders can pursue a collaborative approach by forming international partnerships, pooling resources, and aligning on ethical AI frameworks that promote safe, inclusive, and transparent AI deployment worldwide.

In both the US and China, AI development benefits from considerable investment, access to large volumes of data, and extensive government and corporate support. The US benefits from a robust ecosystem of technology companies, vast financial resources, and a regulatory environment conducive to rapid innovation. In contrast, China's AI growth is fuelled by strategic state support, access to massive datasets, and an ambitious agenda to lead in key AI areas. These factors grant both countries a competitive edge, presenting challenges for Europe, where AI development is often tempered by regulatory constraints and a fragmented market.

European RDI organisations face particular hurdles in securing the funding and resources required to compete at the scale and pace of AI research in the US and China. Limited access to large datasets and less integrated AI ecosystems can make it difficult for European institutions to achieve breakthroughs as rapidly. Additionally, Europe's commitment to stringent data privacy laws - like the GDPR - while essential for protecting citizens' rights, can also slow down innovation in data-intensive AI applications, putting European institutions at a potential disadvantage.

UPHOLDING EUROPEAN ETHICAL STANDARDS IN A COMPETITIVE ENVIRONMENT

A key differentiator for Europe is its commitment to ethical and responsible AI. European RDI leaders prioritise transparency, privacy, and fairness, aiming to create AI systems that align with European values and set a global benchmark for ethical standards. However, this commitment can pose challenges in a competitive global market where ethical standards vary widely. The lack of a global consensus on ethical AI practices often forces European organisations to make difficult trade-offs between staying competitive and complying with high ethical standards. Leaders must navigate these tensions, promoting AI that meets Europe's core values while finding innovative ways to maintain global relevance.

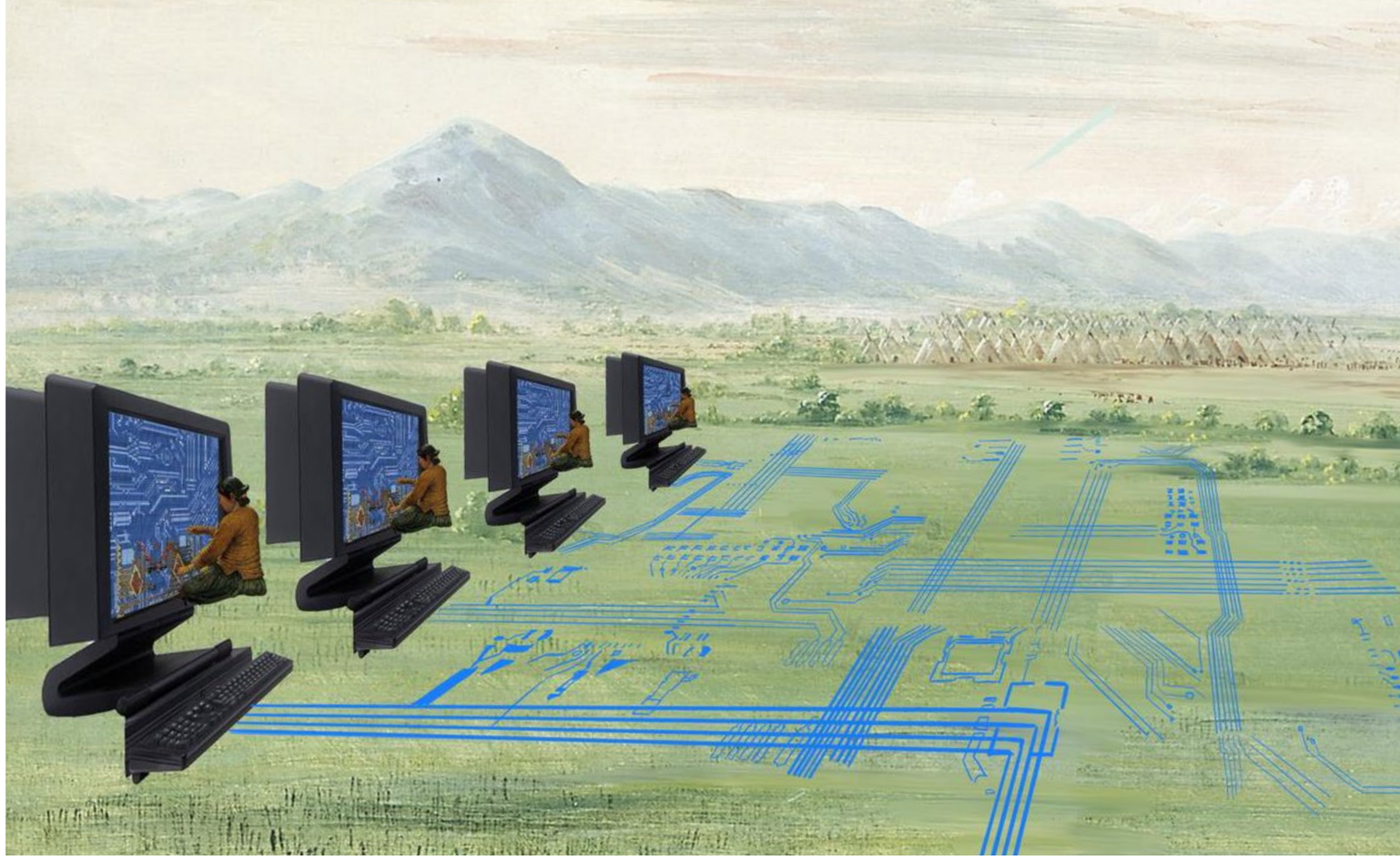
European leaders face the dual challenge of fostering innovation while upholding strict ethical principles. Balancing regulatory compliance with the flexibility needed to explore cutting-edge AI applications is essential. Overly restrictive regulations could stifle innovation, making it challenging for Europe to remain competitive. Leaders must therefore advocate for balanced policies that protect public interests and support ethical AI, while allowing flexibility for innovation.

To address these constraints, European RDI organisations must favour collaboration within Europe and beyond to enhance their influence on the global AI stage. Leaders should focus on cross-border partnerships, pooling resources and knowledge to bridge fragmentation in the European AI ecosystem. By fostering alliances between academic institutions, governments, and the private sector, Europe can strengthen its AI capabilities and create a unified front to compete with concentrated AI powerhouses in the US and China.

Europe's leadership in ethical AI also translates into a unique opportunity to influence global standards. European RDI organisations and leaders can play a pivotal role in establishing guidelines for AI ethics and governance, setting an example that other regions may follow. Engaging in international dialogue on AI policy and regulation allows Europe to promote its vision of responsible AI, potentially shaping global standards and norms. This diplomatic approach can help ensure that Europe's values of transparency, privacy, and human-centred AI integrate international AI frameworks, positioning European leaders as champions of ethical innovation.

To remain competitive, European RDI organisations must strategically invest in AI areas that leverage their unique strengths. Leaders should carefully assess priorities, identifying fields where Europe – and individual countries, like Portugal – can excel, such as AI applications in healthcare, green technology, oceanic resources, tourism, and manufacturing. Focusing on these domains provides Europe with a competitive advantage, aligning innovation efforts with the region's regulatory environment and societal priorities.

Investment in AI infrastructure and talent is also crucial for Europe's long-term competitiveness. Leaders should advocate for sustained funding to support advanced AI research, computational resources, and the development of skilled personnel. Programmes that attract top talent from within Europe and abroad can help bridge the skills gap, fostering the expertise needed to drive AI innovation. Additionally, investing in AI education and training will prepare the next generation of European leaders to address the geopolitical challenges of a rapidly evolving AI landscape, strengthening Europe's role on the global stage.



Hanna Barakat + AIxDESIGN & Archival Images of AI / Better Images of AI / Weaving Wires 2 / CC-BY 4.0

CONCLUSION

The integration of AI into research, development, and innovation organisations presents leaders with complex challenges and transformative opportunities. To succeed, leaders must adopt a vision that extends beyond individual organisational gains, championing ethical AI practices, fostering collaborative human-AI ecosystems, and enabling societal change through a culture of openness and experimentation. Leaders must also be adept at managing change in a volatile environment, guiding their institutions through AI-driven transformation while upholding ethical standards.

European RDI leaders face an additional array of geopolitical challenges, including competition from AI giants like the US and China and the need to balance ethical standards with the demands of rapid innovation. By fostering collaboration, advocating for balanced regulatory frameworks, and strategically investing in AI domains that align with Europe's values, leaders can position Europe as a significant player in AI. Navigating this complex geopolitical environment requires a blend of strategic foresight, ethical commitment, and adaptability.

As RDI organisations move forward, investing in leadership that understands both the technical and ethical dimensions of AI is paramount. By fostering collaboration, championing ethical standards, and promoting a culture of continuous learning and adaptation, leaders can ensure that AI becomes a force for good, advancing both organisational and societal goals. The future of AI in RDI organisations depends on visionary leadership that can balance innovation with responsibility, navigating uncertainty to drive meaningful and sustainable impact.

Effective leadership will require blending strategic foresight, ethical responsibility, and adaptability. Through cross-sector collaboration, upholding ethical standards, and empowering teams to work alongside AI, leaders can ensure that AI contributes to sustainable progress for both their organisations and society.

BUILDING INCLUSIVE AI: INTEGRATING RELATIONAL ETHICS WITH COMPOSITIONAL DESIGN

VIRGINIA DIGNUM
Umeå University
virginia.dignum@umu.se

As AI technologies advance, ensuring their transparency, scalability, and ethical governance becomes increasingly vital. Traditional AI paradigms, rooted in a rational framework [9], prioritize efficiency, problem-solving, and optimization. This paradigm emphasizes autonomy and a task-oriented approach, but often leads to AI systems that are monolithic and opaque. Moreover, this approach neglects crucial aspects such as societal impact, ethical considerations, and long-term maintainability [10]. To address these gaps, we propose integrating relational ethics, exemplified by feminist philosophies, with compositional AI principles from software engineering. This integration emphasizes community, interconnectedness, modularity, and transparency, ensuring that AI systems are both socially responsible and technologically robust, capable of meeting the complex demands of modern socio-ecological systems. That is, by embracing a structured, modular approach and a relational, community-centric ethic, AI can better meet the complex demands of modern socio-ecological systems, ensuring responsible and sustainable technological development.

Current AI systems, particularly Generative AI (GenAI), are designed to prioritize immediate performance over long-term maintainability and ethical considerations. This leads to several issues, including complexity and opacity, as the interconnected nature of large AI models makes them difficult to navigate and understand, leading to challenges in verifiability and governance. Additionally, these AI approaches often reinforce existing power structures and fail to account for societal biases, as they are designed from a rational, individualistic perspective [1]. Furthermore, the societal impact of AI, including its potential to reinforce gender and racial biases, is not adequately addressed in current paradigms [2].

To address these critical limitations, a new paradigm is needed that shifts the focus from rationality to sociality [4]. Feminist epistemology challenges the traditional, individualistic concept of AI by emphasizing empathy, sensitivity, and community. It aims to include principles of accountability, responsibility, and transparency in AI design and use, focusing on how AI systems can affect and be affected by societal values and power structures [1, 6]. Other philosophical schools, such as Ubuntu, a non-Western philosophy rooted in African traditions, also emphasize interconnectedness, community, and the idea that one's humanity is tied to the humanity of others. Such alternative paradigms promote norms of reciprocity, selflessness, and symbiosis, advocating for an AI that enhances communal relationships and is sensitive to the societal context in which it operates and help mitigate biases and promote social justice [5].

At the same time, to address the technical limitations of current AI systems, a shift is needed towards modular AI systems that draw on established software engineering principles such as modularity, abstraction, and separation of concerns [3]. This compositional AI paradigm promises greater flexibility, transparency, and integration with human expertise. Decomposing complex AI systems into smaller, manageable components allows for independent development, testing, and optimization, enhancing flexibility and transparency. Simplifying interfaces and protecting data ensures that AI systems are robust, reliable, and maintainable [10]. Combining AI with human expertise and analog models adds robustness and contextual relevance, especially in scenarios with significant uncertainty or ethical considerations.

By integrating relational ethics with compositional AI principles, we can develop AI systems that are both socially responsible and technologically robust. This integration involves embedding ethical considerations, transparency, and accountability into the core design of AI systems, ensuring they align with human values and societal needs [7]. Furthermore, cross-disciplinary collaboration in the development and deployment of AI systems is essential to address to the complex interactions and dependencies characteristic of modern socio-technicalecological systems [8].

Despite its potential, developing a modular, hybrid, and human-centric AI paradigm presents several significant challenges. Managing interactions between distributed nodes and ensuring data privacy and security require advanced techniques and robust infrastructure. Developing transparent and verifiable AI operations, including explainable AI techniques, is essential for building trust and accountability. To effectively integrate relational ethics with a compositional approach to AI development and use, the following characteristics and properties are essential:

Hanna Barakat + AIxDESIGN & Archival Images of AI / Better Images of AI / Textiles and Tech 2 / CC-BY 4.0

- **Modularity:** AI systems should be decomposed into smaller, manageable components that can be independently developed, tested, and optimized. This enhances flexibility, transparency, and maintainability.
- **Transparency:** Systems should be designed to be transparent in their operations and decision-making processes. This includes the ability to explain decisions in a way that is understandable to humans, thus building trust and accountability.
- **Accountability and Responsibility:** AI systems should include mechanisms to ensure accountability for their actions and decisions. This involves clear documentation of decision-making processes and the ethical considerations involved.
- **Societability:** The design of AI systems should emphasize community values and interconnectedness, ensuring that the systems enhance communal relationships and are sensitive to societal contexts .
- **Ethical Considerations:** Embedding ethical considerations into the core design and operation of AI systems is crucial. This includes addressing biases, promoting fairness, and ensuring that AI benefits all sections of society.
- **Scalability and Maintainability:** Systems should be designed to scale effectively without compromising performance or reliability. This includes ensuring that the systems are maintainable over the long term.
- **Human-Centric Design:** AI systems should be designed with a humancentric approach, ensuring that they augment human capabilities and are aligned with human values and societal needs.
- **Cross-Disciplinary Collaboration:** Developing and deploying AI systems should involve cross-disciplinary collaboration to address diverse challenges and perspectives. This fosters a more comprehensive and inclusive approach to AI development.

Image generated with AI using OpenAI's DALL·E

By integrating these characteristics and properties, we can develop AI systems that are both socially responsible and technologically robust. However, several challenges must be addressed to realize this vision. Ensuring transparency and accountability requires overcoming the complexity of AI systems, which can obscure decision-making processes. Embedding ethical considerations into AI design involves tackling ingrained biases and ensuring fairness across diverse societal contexts. Achieving scalability and maintainability without sacrificing performance demands innovative engineering solutions. Furthermore, fostering cross-disciplinary collaboration requires bridging gaps between different fields and ensuring effective communication.

Despite these challenges, this integrated approach fosters a more inclusive and sustainable technological future, addressing the multifaceted challenges of modern socio-ecological systems and ensuring the responsible and ethical use of AI. Adopting a ethical relational approach to AI, combined with a structured, engineering-focused compositional paradigm, will lead to the development of AI systems that are not only powerful and efficient but also aligned with human values and societal needs. This integrated approach paves the way for a future where AI serves as a force for positive change and inclusivity.

REFERENCES

1. Alison Adam. Artificial intelligence and women's knowledge: What can feminist epistemologies tell us? Women's Studies International Forum, 18(4):407–415, 1995.
2. Abeba Birhane. Algorithmic injustice: a relational ethics approach. Patterns, 2(2):100205, 2021.
3. Virginia Dignum. Social agents: bridging simulation and engineering. Communications of the ACM, 60(11):32–34, 2017.
4. Virginia Dignum. Responsible Artificial Intelligence: Recommendations and Lessons Learned, pages 195–214. Springer International Publishing, Cham, 2023.
5. Mark Dingemanse, Andreas Liesenfeld, Marlou Rasenberg, Saul Albert, Felix K Ameka, Abeba Birhane, et al. Beyond single-mindedness: A figureground reversal for the cognitive sciences. Cognitive Science, 47(1):e13230, 2023.
6. Catherine D'Ignazio. What would feminist data visualization look like. MIT Center for Civic Media, page 20, 2015.
7. Vinodkumar Prabhakaran, Margaret Mitchell, Timnit Gebru, and Iason Gabriel. A human rights-based approach to responsible ai, 2022.
8. Catharina Rudschies, Ingrid Schneider, and Judith Simon. Value pluralism in the ai ethics debate–different actors, different priorities. The International Review of Information Ethics, 29(1):6–23, 2021.
9. Stuart Russell and Peter Norvig. Artificial intelligence: a modern approach. Prentice Hall, 2010.
10. D Sculley, G Holt, D Golovin, E Davydov, T Phillips, D Ebner, et al. Hidden technical debt in machine learning systems. In Advances in Neural Information Processing Systems, 2014.

ARTICLE 3

RESPONSIBLE AI:
BEYOND RESEARCH AND REGULATIONS

PEDRO SALEIRO
Opnova
pedrosaleiro@gmail.com

AI is moving at an incredible pace, and it's accelerating, with trillions of dollars under investment. It's reshaping industries and even the way we work and live our lives. However, as AI becomes more influent, the importance of Responsible AI becomes more obvious.

As researchers, engineers and industry leaders, we need to ensure our AI systems aren't just theoretically good—they need to be proven trustworthy when used in everyday situations. There's no shortage of research on fairness, safety, robustness, explainability, and privacy in AI, but there's still a big gap between research work and real-world practice. With the EU AI Act on the horizon and ongoing research in this field, we must ask ourselves: are we truly prepared to implement these principles effectively?

The EU AI Act, set to become a landmark regulation in AI governance, aims to create standards for developing and using AI systems in Europe. However, there's a potential risk that this regulation, if not implemented effectively, will focus more on legalistic compliance—producing mountains of paperwork rather than ensuring that AI systems are thoroughly and comprehensively tested.

The issue here is not the intent of the regulation but how it's implemented. We've seen similar challenges in other industries where innovation was stifled by excessive regulation that focused more on process than outcomes.

Testing is the only way to ensure that AI behaves reliably in a range of environments, including edge cases that could present significant risks. A good analogy is to look at mission-critical industries like aerospace or nuclear energy, where failure is not an option. In these sectors, thorough testing is built into every stage of the development process, from initial design to final implementation. AI should be no different.

Traditional AI testing methods, which focus on specific datasets and controlled environments, won't be sufficient for modern AI systems, particularly those we classify as "agentic" – systems capable of perceiving their environment, making decisions, delegating and taking actions to achieve specific goals.

Take, for example, a fraud detection agent in a banking system. This agent could request a fraud score from a tabular fraud detection model, visually scan transaction histories, cross-reference credit card data, search for patterns across different devices, and even contact the account holder. Based on all this information, it might decide whether or not to block the account. This type of system is incredibly powerful but also incredibly risky if not properly tested.

In the case of agentic AI, this means testing not just individual tasks but the entire decision-making pipeline. For instance, we need to simulate complex, real-world fraud scenarios, such as coordinated attacks across multiple accounts, to ensure the AI behaves as expected under these high-stakes conditions.

These AI agents, built on large multimodal models, present several unique challenges:

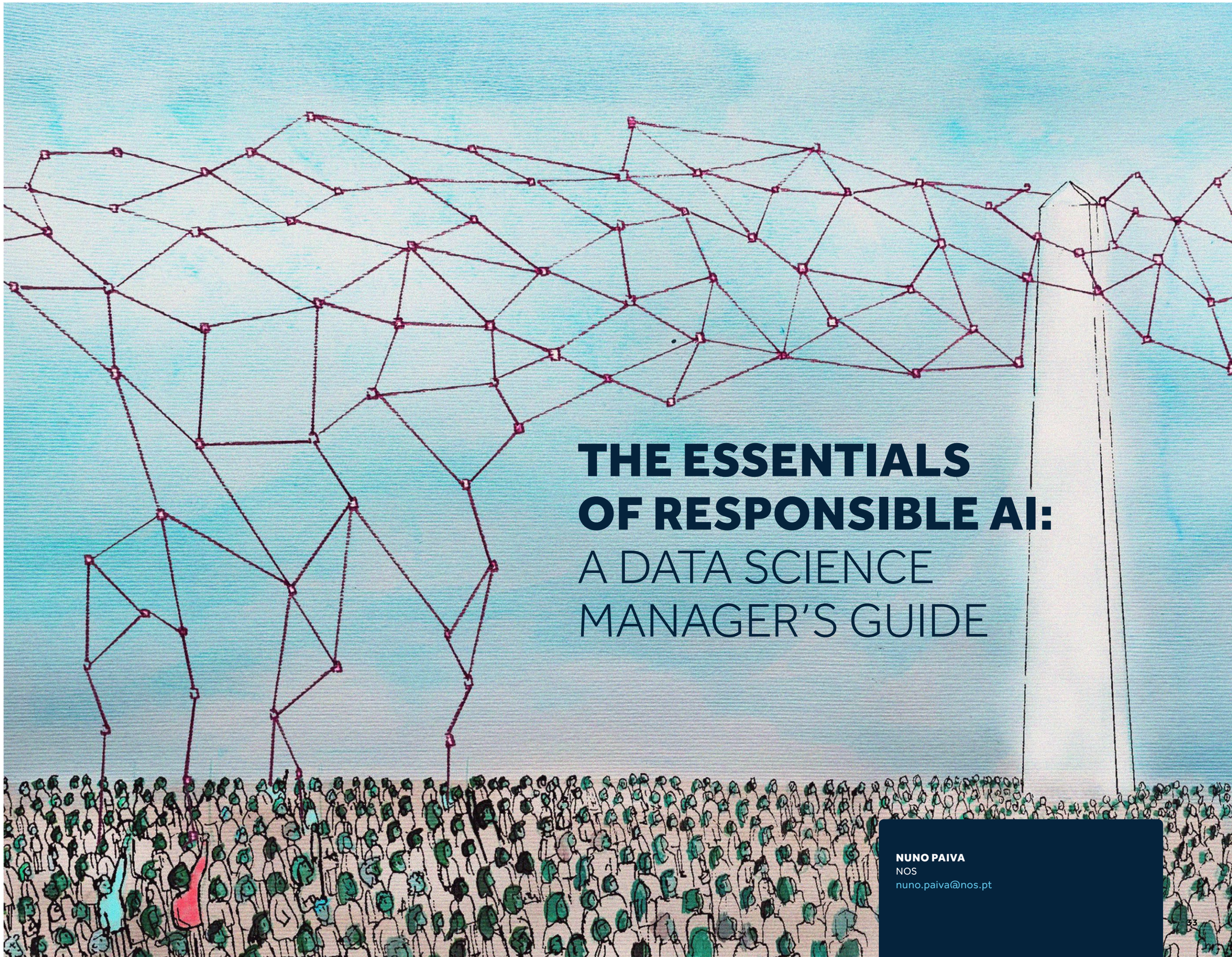
1. **Non-determinism:** Unlike traditional software, agentic AI systems may produce different outputs for the same input, making reproducibility and bug identification more complex.
2. **Non-stationarity:** These systems can learn and adapt over time, potentially changing their behavior in ways that may not be immediately apparent or predictable.
3. **Complexity**: The intricate, multi-component nature of these systems makes it difficult to isolate and test individual parts without considering the whole.
4. **Contextual performance:** The performance of agentic AI can vary significantly based on the context in which it operates, requiring testing across a wide range of scenarios.
5. **Ethical considerations:** As these systems make increasingly consequential decisions, we must test not just for functionality, but also for alignment with human values and ethical principles.

To address these challenges, we need a concerted effort from both academia and industry to develop comprehensive, standardized open-source AI testing frameworks that should enable:

1. **statistical testing:** Given the non-deterministic nature of these systems, we need to move beyond simple input-output testing to statistical approaches that can quantify behavior across distributions of outcomes.
2. **continuous testing:** As AI agents learn and evolve, testing must be an ongoing process throughout the AI lifecycle, from development to deployment and beyond.
3. **multi-component testing:** Frameworks should allow for testing of individual components as well as the system as a whole, helping to isolate issues and understand complex interactions.
4. **ethical evaluation:** Beyond functional testing, we need methodologies to assess the ethical implications of AI decisions and behaviors.
5. **scenario-based testing:** Tools should support the creation and execution of diverse, realistic scenarios to evaluate AI performance across different contexts.

For the EU AI Act to really work, we have to go beyond just meeting the basic legal requirements. Instead, we need to create a culture of continuous and serious testing that keeps up with AI's fast pace. This is not just about following rules. These tools will not only help us create more reliable and trustworthy AI systems but will also accelerate the pace of innovation by providing developers with the confidence to push the boundaries of what's possible.

It's time for European researchers, engineers and industry leaders to walk the talk. Responsible AI requires more than regulation—it requires a commitment to continuous testing and improvement, so that we can build a future where AI enhances human capabilities while safeguarding human rights and the EU societal values.



THE ESSENTIALS OF RESPONSIBLE AI:
A DATA SCIENCE MANAGER'S GUIDE

NUNO PAIVA
NOS
nuno.paiva@nos.pt

Jamillah Knowles & We and AI / Better Images of AI / People and Ivory Tower AI / CC-BY 4.0

WHY DOES RESPONSIBLE AI (RAI) MATTER?

"The most important thing is not life, but the good life." This quote, attributed to Socrates, resonates profoundly in today's rapidly evolving technological landscape, where our pursuit of a good moral life is increasingly intertwined with the tools we develop and use. Technology is not ethically neutral; it significantly shapes our values, behaviours, and societal norms.

AI can make decisions that impact our lives, from recommending products to making healthcare or hiring decisions. This is why organisations must consider ethics—not just their business goals—when developing AI systems. Yet, many companies struggle to balance their ethical values with their day-to-day practices.

The AI Incident Database[1] highlights the challenges posed by AI systems by tracking instances where they have caused harm or near-harm. In 2023, 123 incidents were recorded, marking a 32% increase compared to 2022 – with a constant rise rise in reported cases over recent years. While high-stakes applications, such as the predictive AI tools developed during the COVID-19 pandemic for patient diagnosis and triage[3], exemplify the potential consequences of poorly implemented AI systems, these concerns are not limited to critical situations. For instance, the case of Staples[2], which varied online prices based on user location and demographics, caused reputational damage and illustrates how AI-driven practices can lead to significant public backlash. Notably, this incident has not been reported in the AI Incident Database, underscoring the issue of underreporting. This upward trend highlights the urgent need for organisations to prioritise RAI practices to mitigate risks and prevent further issues.

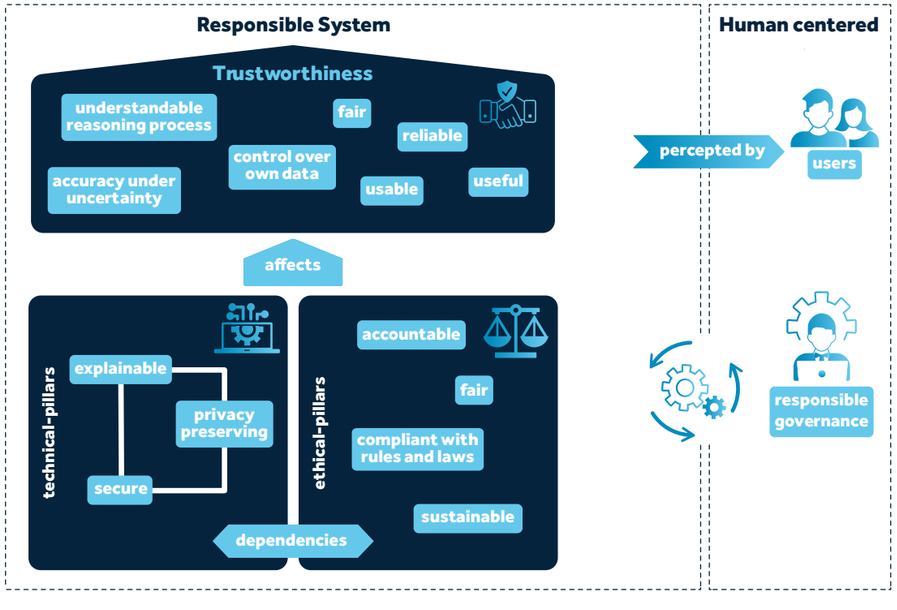
However, defining what constitutes RAI is not straightforward. Recent research[4] points out the challenges of inconsistent terminology and overlapping concepts like "trustworthy AI" and "ethical AI." These terms are often used interchangeably, creating confusion, and making it harder for organisations to understand and implement RAI principles effectively. Simply promoting trust in AI systems is insufficient. To truly matter, RAI must be rooted in ethical considerations that align with societal values and legal norms.

A clearer definition of RAI, proposed through an analysis of 254 papers, emphasises a human-centred approach—prioritising the well-being, rights, and needs of individuals affected by AI systems. This approach ensures user trust by promoting ethical decision-making that is fair, accountable, and consistent with societal laws and norms. It also includes sustainability, ensuring that AI systems consider long-term societal and environmental impacts. Additionally, RAI ensures that automated decisions are explainable to users and that their privacy is preserved through secure implementations.

WHAT LIES BENEATH RESPONSIBLE AI?

While the importance of RAI is clear, it is underpinned by a robust framework designed to integrate both technical and ethical considerations. This research-proposed framework [4], visually represented in Figure 1, emphasises the interdependence of the technical and ethical pillars of RAI. Together, these pillars are managed through responsible governance, resulting in systems perceived by stakeholders as trustworthy. To illustrate what's at stake for each pillar, we will use a hiring system as the conducting use case, demonstrating how these principles apply in practice.

Figure 1 Research proposed RAI Framework



TECHNICAL PILLARS

- **Explainability:** The hiring system should provide clear reasons for why it selected or rejected candidates. For example, it could highlight that a rejected candidate lacked specific skills listed in the job description. This transparency ensures the decision is understandable and helps build trust among applicants and employers.
- **Privacy-Preserving:** The system must ensure the privacy of applicants by securely handling sensitive personal data like names, addresses, and employment history. For instance, it can anonymise candidate data during analysis, ensuring that employers don't have access to identifiable information until later in the process.
- **Security:** The hiring system must be designed to resist data breaches, ensuring that applicant information is protected from unauthorised access. For example, the system could use encryption to protect personal details such as social security numbers.

ETHICAL PILLARS

- **Fairness:** The system must avoid biases such as ones related to race, gender, or socioeconomic background. For example, the algorithm should be audited to ensure that it doesn't favour applicants from specific educational institutions or demographics, maintaining equal opportunities for all.
- **Accountability:** The development team is responsible for monitoring the system to correct any discriminatory practices that arise. For instance, if they detect a bias pattern, the developers must intervene, adjust the algorithm, and report on the issue.
- **Compliance:** The system, depending on the sector, should comply with laws, such as in the case of hiring anti-discrimination regulations (e.g., the U.S. Equal Employment Opportunity Act). Compliance could include regular checks to ensure that hiring recommendations meet legal standards and don't inadvertently violate them.
- **Sustainability:** The system could be designed with long-term fairness in mind, constantly refining its decision-making to account for changing labour markets and societal norms. Additionally, it should be energy-efficient, considering the environmental impact of computational needs.

TRUSTWORTHINESS AND HUMAN-CENTRED DESIGN

- At the core of this framework, trust is built when users perceive the system as fair, transparent, and reliable. If the AI hiring system provides understandable results, ensures privacy, and complies with legal standards, employers and candidates will trust its decisions. By keeping a human-in-the-loop, hiring entities can intervene in cases where the system might fall short—ensuring that human values are respected at all stages.

The RAI framework ensures that the technology not only operates adequately but does so ethically and safely. It integrates technical and ethical pillars to create systems that are explainable, secure, fair, and compliant with regulations. The use of real-world processes, such as in a hiring system, highlights how these pillars work together to foster trust and accountability, illustrating that RAI is not just about technology—it's about its impact on people and society.

WHICH ARE THE ESSENTIAL RAI GOVERNANCE FRAMEWORKS?

As AI continues to evolve, governments, companies, and researchers are developing frameworks to ensure the responsible use of AI systems. Here are some key types of documents related to RAI governance:

- **AI Laws and Regulations:** The global landscape of AI governance is rapidly evolving, with many countries developing legal frameworks to ensure the ethical use of AI. In the U.S., initiatives like the SAFE Innovation Framework aim to balance AI development with safety. The EU's AI Act takes a risk-based approach, enforcing strict rules for high-risk AI applications. According to the Global AI Legislation Tracker[5], other nations are also advancing their own AI regulations, reflecting a growing recognition of the importance of AI governance, with each framework tailored to its unique social, legal, and economic context.
- **Guidance and Frameworks:** Various guidance documents and frameworks help organisations implement RAI practices. The NIST AI Risk Management Framework[6] provides guidance on identifying and managing AI risks, with a focus on ethics and adaptability. The Turing Institute's Ethics and Safety Framework[7] emphasises fairness, accountability, and transparency to ensure safe and ethical AI systems. The OECD AI Principles[8] provide a global benchmark for ethical AI, focusing on transparency, robustness, and inclusivity. In addition to these, companies like Microsoft publicly share their AI Principles[9] - focusing on fairness, reliability, and privacy - demonstrating the industry's dedication to responsible and ethical AI governance.
- **Standards and Certifications:** Standards like ISO/IEC 42001[10] advance guidelines for RAI, focusing on transparency, accountability, and ethical practices to build trust in AI systems. These standards allow organisations to assess their AI systems against measurable benchmarks. Additionally, there are now several various RAI Certification Programs[11] helping organisations demonstrate ethical AI practices. While these certifications are voluntary and lack the regulatory force of the EU AI Act, they play a critical role in promoting a culture of RAI across industries.
- **National AI Strategies:** National AI strategies serve as comprehensive frameworks that integrate AI laws, regulations, and existing standards. For instance, Canada's Pan-Canadian Artificial Intelligence Strategy focuses on three pillars: Commercialization, Standards, and Talent and Research. The Standards Council of Canada develops standards like CAN/CIOSC 101:2019 [12], which establishes minimum requirements for the ethical design and use of automated decision systems. This standard aligns with the OECD AI Principles [8] by addressing critical aspects like transparency, accountability, and respect for human rights, reinforcing Canada's commitment to RAI practices. Countries integrating frameworks like the OECD AI Principles and the EU AI Act into their national strategies can ensure that AI development aligns with ethical standards while promoting innovation and economic growth.

Image generated with AI using OpenAI's DALL·E

Given the diversity of RAI governance frameworks, organisations must acknowledge that a "one-size-fits-all" approach is inadequate. To effectively implement RAI, organisations need to assess their specific contexts, capabilities, and objectives. This assessment is crucial for selecting and adapting frameworks that align with their ethical commitments and operational realities.

For example, a start-up developing an AI-driven hiring platform may prioritise implementing the OECD AI Principles to ensure fairness, transparency, and accountability in its algorithms. By focusing on these ethical design principles, the start-up can build trust with users and differentiate itself in the marketplace.

In contrast, a multinational corporation using the start-up's hiring system must navigate more complex compliance challenges. This organisation is likely to emphasise compliance with the EU AI Act, which mandates thorough accountability and risk management for high-risk AI applications. The multinational's concerns may include ensuring robust reporting, bias detection, and comprehensive data governance to mitigate reputational risks and comply with stringent regulations.

Thus, while both the start-up and the multinational are engaged in algorithmic hiring, their different priorities lead them to adopt distinct RAI frameworks. The start-up focuses on ethical design principles to foster trust, whereas the multinational favours compliance and risk management to safeguard operations. This example illustrates the importance of tailoring RAI frameworks to meet the unique needs and challenges of various organisations.

HOW MATURE IS YOUR RAI APPROACH?

Over the past four years, McKinsey, a consultancy company, has identified "AI High Performers"[13], companies that derive over 20% of their EBIT from AI, excelling in strategy, talent, and technology. Their 2022 Digital Trust survey [14] shows a strong correlation between RAI practices and business performance; organisations prioritising digital trust see annual revenue and EBIT growth of at least 10% more often than their peers.

Although we don't have research to support the notion that EBIT drives RAI programmes or vice versa, studies suggest that organisations demonstrating strong performance also tend to be more prepared, with faster and more robust practices. Whether it is RAI pushing EBIT or the other way around, these organisations are clearly better positioned for success. For example, 70% of leaders in digital trust have adopted automated models that prevent failures, compared to less than 40% of others. Additionally, the increasing focus on fairness, explainability, and security demands new skills from AI teams developing greater expertise in validating, debugging and discover new knowledge while deploying models.

Despite progress among top performers, advancements in addressing AI risks have been slow across the industry. According to McKinsey's State of AI 2022 survey[15] , only 22% of companies adequately explain their AI decision-making processes.

A survey by BCG [16], another leading consulting firm, highlights a gap between intention and execution—while 42% of organisations perceive AI as a strategic priority, only 19% have fully implemented RAI programmes. Among the 16% of "RAI Leaders," RAI is integrated into their broader corporate responsibility strategies, which increasingly align with sustainable development goals, such as reducing carbon footprints and promoting environmental stewardship.

A relevant example in the telecommunications sector is NOS Comunicações, a leading company in Portugal. Recently, the organisation undertook a project to enhance mobile network performance by deactivating network elements during periods of low usage, resorting to customer data to optimise this process while maintaining service quality. While this initiative is a commendable way to comply with SDGs, representing a commitment to RAI principles, it is also a less contentious use case - since it leads to cost savings for the company through improved operational efficiency and reduced energy consumption. Nonetheless, prioritisation and effective execution in said projects contribute significantly to building a culture of responsibility and innovation, reinforcing the importance of integrating RAI into the core business strategy.

This integration is further addressed in key survey findings that show significant differences between RAI Leaders and other organisations:

- 74% of RAI Leaders say RAI is on their top management agenda, compared to 46% of non-Leaders.
- 77% of Leaders are willing to invest in RAI initiatives, versus 39% of non-Leaders.
- 73% of Leaders connect RAI to corporate social responsibility, compared to 35% of non-Leaders.

Analogous to Tom Davenport's work on analytics [17], where the adoption of advanced analytics capabilities provided companies with a competitive edge, today's leading organisations are similarly gaining advantages by embedding RAI into their core values. Just as analytics leaders thrived by prioritising data-driven decision-making, those that emphasise RAI are positioning themselves for sustainable long-term success.

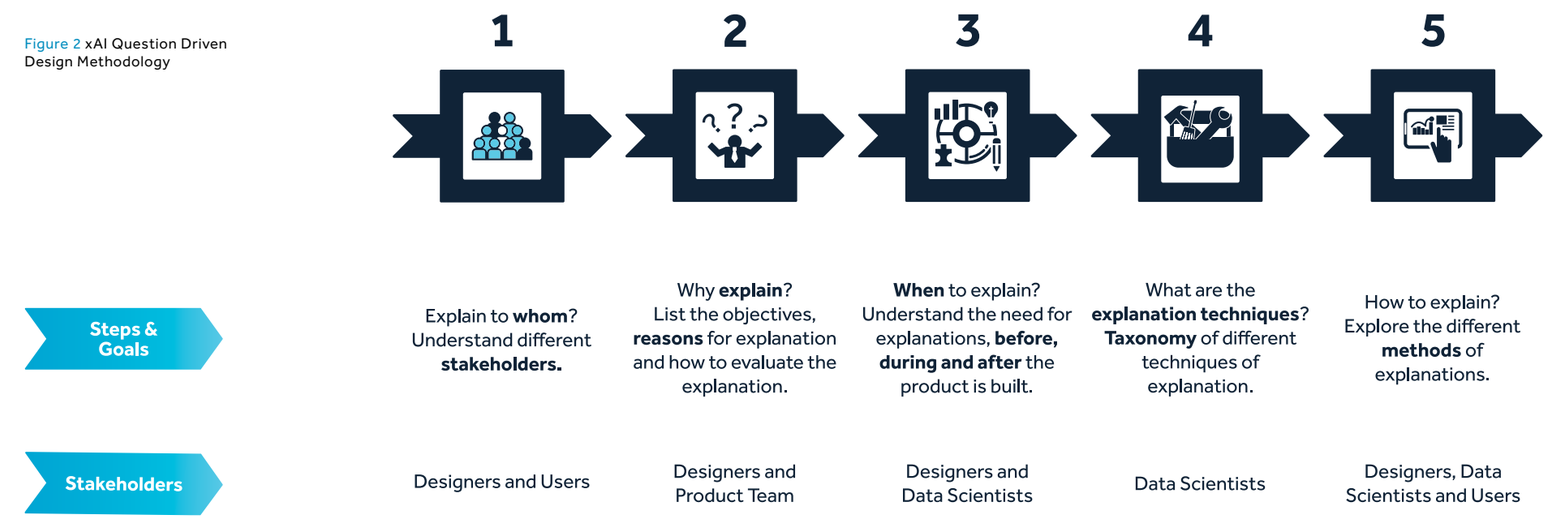
ESSENTIAL RECOMMENDATIONS FOR ASPIRING RAI LEADERS: INSIGHTS FROM A DATA SCIENCE MANAGER

BCG's survey also identifies which are the organisational constraints that limit organisations' ability to implement RAI initiatives, which we can group into two main reasons:

- **Talent Development:** Two challenges highlighted are the lack of RAI expertise and insufficient training among staff. To integrate RAI into every AI project, the workforce must develop general skills, including AI literacy, which encompasses understanding AI concepts, identifying potential harms, and implementing inclusive AI practices.
- **Senior Engagement:** The report indicates that two of the five major barriers to RAI adoption stem from senior leadership. A lack of prioritisation, attention, and resources hampers the strategic focus necessary for advancing RAI initiatives.

Despite these challenges, there is always potential for progress, particularly through the proactive role of data science managers. In my experience, I have seen how AI initiatives can transform organisations; equally important is fostering RAI practices that meet business goals and ethical standards. As leaders, we must set the tone by demonstrating commitment to RAI, being transparent about ethical challenges, and involving our teams in crafting solutions. This approach cultivates accountability and shared responsibility.

Figure 2 xAI Question Driven Design Methodology



HERE ARE ACTIONABLE RECOMMENDATIONS FOR ASPIRING RAI LEADERS:

1. Focus on Training and Upskilling

Technical teams in AI/ML must acknowledge that their decisions can significantly impact society. Although organisations may take time to fully establish RAI practices, early preparation is essential to accelerate this journey. Training is crucial for raising awareness; however, quality hands-on RAI education is limited. For instance, research [18] found that only 22 of 186 machine learning courses at leading U.S. universities included ethics-related content, exposing a gap in technical training.

To address this, leveraging existing research [19], we developed a hands-on course aimed at improving the ethical use of AI in decision-making, focusing on enhancing AI explainability. Identifying the challenges in providing users with clear information, we blend techniques from explainable AI (XAI) to inform user experience (UX) design. The course introduces a Question-Driven Design Process that aligns user needs with the selection and implementation of XAI techniques, fostering collaboration between designers and AI engineers. Through practical applications, participants learn to effectively tackle the design challenges of AI systems following a methodology as described in Figure 2, and using a practical loan application case study. For instance, designing explanations for a risk manager stakeholder that aims to compare similar loan applications requires a contrastive explainability method, while for the end user (customer), counterfactual explanations could help to provide actionable insights to improve likelihood of loan approval – e.g.,: how much a clean default history can change the loan decision process. The course is now available as an elective in a partner university's master's programme, expanding its impact on future professionals.

2. Applying RAI Principles to Real-World Projects

The next step was to identify projects for applying RAI principles. While fairness often appears in clear use cases like loan approvals or hiring, where unbiased decisions across sensitive attributes (e.g., gender, age) are critical, real-world ethical concerns can be less straightforward.

Understanding that fairness is vital for sustainable business outcomes requires it to be framed appropriately within the business context. For example, in a call-centre project, our initial focus on short-term optimisation led to experienced operators handling more calls, creating an imbalance. By reframing the challenge as an opportunity to train operators through balanced call assignments, we achieved a fairer workload distribution. This not only addresses fairness but also fosters long-term workforce sustainability and improves operator well-being.

3. Fostering Team Collaboration for RAI

RAI initiatives must extend beyond data science and AI teams to include legal, compliance, and corporate social responsibility (CSR) departments. Engaging these teams ensures balance between AI projects and broader company policies, creating opportunities to enhance processes toward RAI-centric practices.

For instance, during project initiation, we carry out a thorough risk assessment in collaboration with the compliance team. This assessment can be improved by integrating RAI-related technical knowledge. A key question to consider is, "How do you justify the model's complexity for the specific use case?"

While this question may seem straightforward, a lack of understanding regarding glassbox and blackbox models can lead to vague justifications. This may result in deploying overly complex models, wasting time on post-hoc explanations that can be misleading[20].

By collaboratively addressing these questions, teams foster shared knowledge and responsibility, ultimately improving RAI outcomes.

As managers, we must bridge the gap between technical teams and senior leadership, advocating for resources and commitment to RAI. Ethical AI is not just a "nice-to-have"—it's essential for reducing risk, improving trust, and aligning with long-term business goals.

By demonstrating how RAI practices like fairness and transparency drive both ethical outcomes and business success, we ensure AI systems benefit not just the company but also society. Leading with this mindset positions our organisations as RAI leaders, ensuring sustainable growth and positive societal impact.

Clarote & AI4Media / Better Images of AI / User/Chimera / CC-BY 4.0

REFERENCES

1. R. A. I. Collaborative, "Artificial Intelligence Incident database." [Online]. Available: https://incidentdatabase.ai/.
2. J. Valentino-DeVries, J. Singer-Vine, and A. Soltani, "Websites Vary Prices, Deals Based on Users' Information." 2012, [Online]. Available: https://www.wsj.com/articles/SB10001424127887323777204578189391813881534.
3. W. D. Heaven, "Hundreds of AI tools have been built to catch covid. None of them helped." 2021, [Online]. Available: https://www.technologyreview.com/2021/07/30/1030329/machine-learning-ai-failed-covid-hospital-diagnosis-pandemic/.
4. S. Goellner, M. Tropmann-Frick, and B. Brumen, "Responsible Artificial Intelligence: A Structured Literature Review." 2024, [Online]. Available: https://arxiv.org/abs/2403.06910.
5. I. A. of Privacy Professionals, "Global AI Law and Policy Tracker." 2024.
6. N. A. Team, "NIST AIRC - PlayBook." [Online]. Available: https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook.
7. D. Leslie, "Understanding artificial intelligence ethics and safety," CoRR, vol. abs/1906.0, 2019, [Online]. Available: http://arxiv.org/abs/1906.05684.
8. "How countries are implementing the OECD Principles for Trustworthy AI - OECD. AI." [Online]. Available: https://oecd.ai/en/wonk/national-policies-2.
9. "Responsible AI Principles and Approach | Microsoft AI." [Online]. Available: https://www.microsoft.com/en-us/ai/principles-and-approach.
10. ISO, "ISO/IEC 42001:2023." 2023, [Online]. Available: https://www.iso.org/standard/81230.html.
11. IEEE.org, "IEEE CERTIFAIED – The Mark of AI Ethics." [Online]. Available: https://engagestandards.ieee.org/ieeecertifaied.html.
12. "CAN/CIOSC 101:2019 (R2021) Ethical Design and Use of Automated Decision Systems." 2019, [Online]. Available: https://scc-ccn.ca/standardsdb/standards/4029998.
13. "The state of AI in 2020." 2020, [Online]. Available: https://www.mckinsey.com/capabilities/quantumblack/our-insights/global-survey-the-state-of-ai-in-2020.
14. "Why digital trust truly matters." 2022, [Online]. Available: https://www.mckinsey.com/capabilities/quantumblack/our-insights/why-digital-trust-truly-matters.
15. "The state of AI in 2022—and a half decade in review." 2022, [Online]. Available: https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2022-and-a-half-decade-in-review.
16. E. M. Renieris, D. Kiron, and S. Mills, "To be a responsible AI leader, focus on being responsible." 2022, [Online]. Available: https://sloanreview.mit.edu/projects/to-be-a-responsible-ai-leader-focus-on-being-responsible/.
17. T. H. Davenport and J. G. Harris, Competing on Analytics: The New Science of Winning. Harvard Business School Press, 2007.
18. J. Saltz et al., "Integrating Ethics within Machine Learning Courses," ACM Trans. Comput. Educ., vol. 19, pp. 1–26, 2019, doi: 10.1145/3341164.
19. Q. V. Liao, M. Pribić, J. Han, S. Miller, and D. Sow, "Question-Driven Design Process for Explainable AI User Experiences." 2021.
20. C. Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead." 2019.

THE FIVE PILLARS FOR COMPANIES TO GET THE MOST OUT OF AI

Luke Conroy and Anne Fehres & AI4Media / Better Images of AI / Models Built From Fossils / CC-BY 4.0

PEDRO AMORIM (1, 2, 3)
pedro.amorim@inesctec.pt

GONÇALO FIGUEIRA (1, 2)
goncalo.figueira@inesctec.pt

(1) Institute for Systems and Computer Engineering, Technology and Science (INESC TEC)
(2) Faculty of Engineering of University of Porto (FEUP)
(3) LTPlabs

The popularity of Artificial Intelligence (AI) has been growing rapidly, due to multiple scientific and technological breakthroughs, as well as their potential application to several areas. The most recent disruption was in generative AI, with large language models (e.g., ChatGPT[1]) accumulating vast amounts of knowledge, and using it to generate text and images in response to human prompts. Those models can be used in different applications, from information retrieval to content generation. However, ChatGPT does not solve all the problems in the world, like those that came before did not. From our experience in interacting with managers across different business sectors, all the glare that came with the emergence of ChatGPT is often undermining the basics of employing AI to improve and transform business processes.

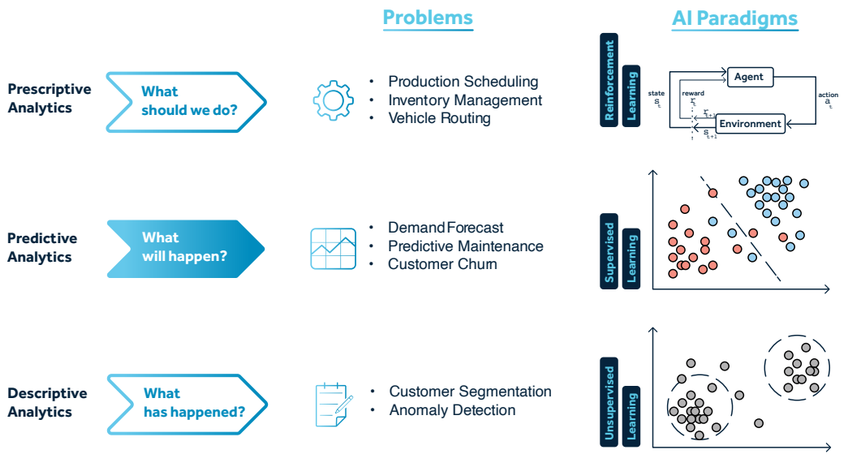
In this article, we review the five main pillars that managers ought to know to be able to convert the promise of AI into reality. These pillars are linked to the need to properly match the existing tools to the tasks to be tackled – this attention is particularly relevant for prescriptive tasks (Pillar #1); the importance of combining different AI approaches to fully address the complex challenges of organisations (Pillar #2); the relevance of focusing on AI methods that are explainable to promote trustworthiness and collaboration (Pillar #3); the possibilities that different human-machine interaction modes open (Pillar #4); and the fundamental activity of ensuring basic AI knowledge across the whole organisation (Pillar #5).

PILLAR #1
MAKE SURE THAT PRESCRIPTIVE TASKS ARE BENEFITING FROM AI

AI has been increasing in companies, supporting decision-making at different levels, especially with descriptive and predictive tasks. Descriptive methods, such as clustering and association rules, allow for instance to segment customers or detect consumer patterns. Predictive methods, such as those used in supervised learning, as the name suggests, can provide predictions, such as sales forecasts or customer churn. However, those methods should not be directly applied to prescriptive problems – problems in which the decision to be made is the core issue to be tackled, such as deciding the best route for a vehicle or the assortment to keep in a store. What we have often witnessed in practice is that managers want to use descriptive and predictive tools for problems that require a different approach. As Abraham Maslow once said, "if the only tool you have is a hammer, you tend to see every problem as a nail".

Prescriptive problems require a different family of AI methods: the likes of reinforcement learning. These methods were in the news almost ten years ago, when AlphaGo[2] was able to defeat the world champion in the ancient game of Go. These are the same methods used in autonomous driving and robot navigation, which are still emerging, due to the complexity and issues involved. Nevertheless, there is an enormous untapped potential for those methods to step into many prescriptive applications, such as order allocation in online retail[3], dynamic scheduling in manufacturing, and dynamic routing[4] in internal or external logistics.

Figure 1 The three levels of Analytics and the natural corresponding AI paradigms.



1 https://chatgpt.com/

2 https://deepmind.google/technologies/alphago/

3 https://www.inesctec.pt/pt/projetos/driven

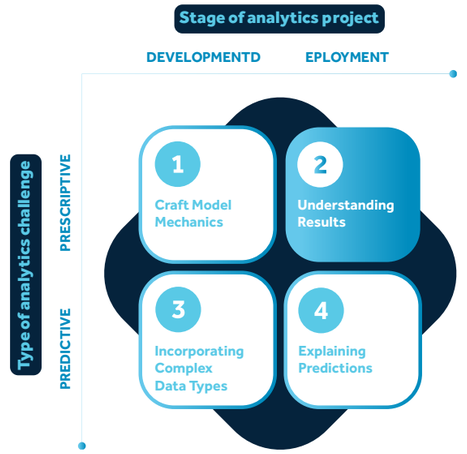
4 Neves-Moreira, F., & Amorim, P. (2024). Learning efficient in-store picking strategies to reduce customer encounters in omnichannel retail. International Journal of Production Economics, 267, 109074.

PILLAR #2
COMBINE MULTIPLE AI METHODS

With the new waves of AI methods emerging from time to time (for example, some years ago it was machine learning, now we have generative AI), it is easy (and lazy!) to think about these technologies as replacements. An equivalent to having the new revamped version of a car substituting its predecessors. Although tempting, this analogy is not accurate. Of course, there are methodological evolutions that render previous algorithms obsolete, but often what we end up having are new tools that can be used in concert to solve ever more complex problems that emerge across society. Consequently, instead of thinking of using the car versions analogy, one can use a *LEGO* analogy in which new blocks can be added to build more interesting creations.

Let us take the example of large language models – known for their natural language capabilities – which can be combined with more traditional predictive (machine learning) and prescriptive (optimisation) algorithms to unlock some of the current challenges with these methodologies[5]. We see opportunities for generative AI to tackle challenges within advanced analytics throughout the development and deployment phases. Large language models can be particularly useful in helping users incorporate unstructured data sources into analyses, translate business problems into analytical models, and understand and explain models' results. This last potential synergy between large language models and advanced analytics is connected to our next pillar.

Figure 2 The four hypotheses to combine generative AI and advanced analytics.



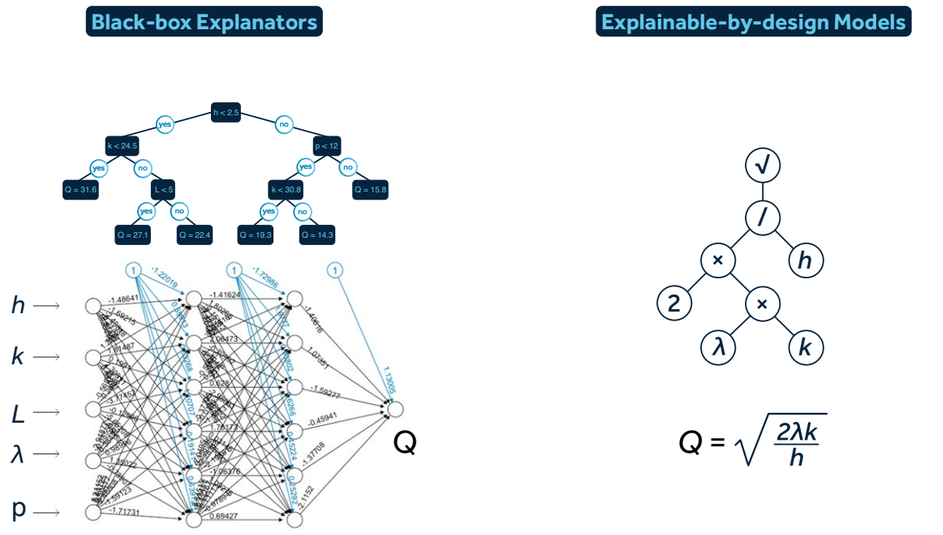
5 Amorim, P., & Alves, J. (2024). How Generative AI Can Support Advanced Analytics Practice. MIT Sloan Management Review, Magazine Summer 2024 Issue

PILLAR #3
OBTAIN EXPLAINABLE AI MODELS

AI models, although performing well in certain tasks, tend to be hard to trust, due to their complexity and, hence, lack of interpretability. This is a major issue hindering adoption in many critical applications, in areas like healthcare, finance, and operations. Explainable AI (XAI) is a growing field, proliferating with multiple research streams. Some propose to use local explanators on top of the AI black-box models (e.g., a decision tree mimicking a neural network). An alternative is to involve human intelligence in the discovery process, resulting in AI and humans working together, in a human-centred, 'guided empirical' learning process. This is made possible by employing 'explainable-by-design' symbolic models and learning algorithms[6].

Symbolic models can be learned by Genetic Programming algorithms. The idea is often to learn a compact model, whose performance does not result from its complexity and fine-tuning of several constants, but rather from learning the structure of the model, by freely combining the problem features with user-defined operators (e.g., arithmetic, logical, etc.). The final model is compact and inspectable and thus does not require an explanator; it is explainable-by-design. In some cases, such a model can have an even higher performance, as it might generalise better. In other cases, the performance may not reach the same level as that of a black-box. In that case, it can be beneficial to keep the black-box and add an explanator on top, to obtain insights into its inner workings.

Figure 3 Example of a black-box model (neural network) and an explainable-by-design model (symbolic expression) for the Economic Order Quantity.



6 The TRUST-AI project (https://trustai.eu/) is doing precisely that.

PILLAR #4
FACILITATE HUMAN-MACHINE INTERACTION

Having humans involved in the discovery process of decision models implies a deep human-machine interaction. However, that level of interaction is not always feasible, e.g., in settings where decision speed is paramount. Also, humans can interact in different ways, such as at the decision level. Decision support systems are exactly that: systems based on advanced methods, suggesting decisions that may or may not be adopted by decision-makers. In some cases, it is important that every decision is evaluated by the human decision-maker, such as when suggesting medical procedures. In other cases, such as fraud detection in credit card transactions, having human agents evaluating all of the thousands of daily predictions is not economically viable.

Despite these constraints, and for companies to get the most out of AI, it is fundamental to grasp that there is no silver-bullet type of solution to set an ideal human-machine interaction. On the contrary, research has identified that companies that are able to be more versatile in the configuration of these interactions are the ones poised to reap more benefits[7]. Sometimes AI decides and implements; other times AI decides and the human implements; but there are variants, such as AI recommends and human decides, or AI generates insights that the human uses in a decision process, or even the human generates solutions, and the AI evaluates. Figuring out the right model can only be (probably) done by humans!

PILLAR #5
ENSURE BASIC AI KNOWLEDGE IN THE ORGANISATION

To be able to manoeuvre a company along the four pillars described above and get the most out of AI, it is key that the organization, from top to bottom, understands the basics of these technologies. Day in and day out we have interactions with managers from different sectors that demonstrate how far from understanding the basics of AI they are. This is, of course, a major roadblock to making good decisions about the use of these technologies, to either improve current business processes or find opportunities to revamp existing business models. As we have heard someone rightfully saying: "If you throw technology (AI) into an 'outdated organisation', the only thing you get is an 'expensive outdated organisation'".

To revert this situation, companies must invest heavily in education initiatives that go up and down the ranks. Interestingly enough, according to Gartner[8], this year, 40% of all organisations will offer or sponsor specialised data science education to accelerate upskilling initiatives. This is a 35 percentage points jump from what we have witnessed in 2021. This must be a continuous effort at the corporate and individual levels, as the pace of evolution of these technologies shows no signs of slowing down.

_____

We believe that these five pillars can be decisive in getting the most out of AI, preventing unnecessary frustrations and, in some cases, making it viable. Managers should print the description of these five pillars and put them up on the wall of their offices, to avoid following the herd and making costly, unimpactful decisions.

7 Ransbotham, S., Khodabandeh, S., Kiron, D., Candelon, F., Chu, M., & LaFountain, B. (2020). Expanding AI's impact with organizational learning. MIT Sloan Management Review

8 https://www.gartner.com/en/data-analytics/insights/data-analytics-skills-competencies

USING CHATGPT IN EDUCATION — A PERSONAL EXPERIENCE

JOSÉ NUNO OLIVEIRA
Institute for Systems and Computer Engineering, Technology and Science (INESC TEC)
School of Engineering, University of Minho
jose.n.oliveira@inesctec.pt

Over the past four years, several events haveshaken the world from the relative tranquillity of the post-World War II period. Towards the end of 2019, a pandemic emerged, the likes of which had not been seen since the 1918 flu. In 2022 and 2023, two wars began on Europe's doorstep; and on the last day of November 2022, the news came that Artificial Intelligence (AI) software capable of automatically carrying out tasks that until then were thought to be the privilege of the human intellect (e.g., writing prose, composing poetry and programming computers) was available online.

It is often said that nothing stays the same after a war or a pandemic. In effect, COVID-19 had a profound impact on education - demonstrating, to the great consternation of some, that traditional face-to-face teaching was becoming obsolete. After all, it was possible to study online, saving time and resources, by learning directly from teaching materials made available by increasingly sophisticated teachers and YouTubers.

As if this weren't enough, the emergence of ChatGPT seemed to appear as yet another disruptive element for teachers, academics and other professionals. Of course, the scientific advances that made large language models (LLM) like ChatGPT possible were well known to experts, researchers and labs, especially to those who developed them. It was also common to use online tools like "Google translate" and others to check a manual translation or even do it automatically in full, and then improve it by hand. Furthermore, people were already used to interactive "bots" showing up here and there in online services. What was awe-inspiring was the broad spectrum of knowledge that ChatGPT seemed to encompass and its conversational fluency. It was no longer just taxi drivers and bank employees who had their jobs at risk: the professional future of educators and researchers themselves, whether academic or not, was threatened.

Many immediately thought of a prohibitionist stance. Others just ignored such services due to their unreliability. In fact, in April 2023, the author of these lines received a message from an Irish colleague about ChatGPT reporting his death (the author's) in 2019. Many anecdotes circulated and still circulate on social media to the delight of humans who ridicule the imperfect engine that tries to replace them. On a different register, "The False Promise of ChatGPT", a guest essay by Noam Chomsky and other linguists on the *NewYork Times* (March 8, 2023) was quite clear about the essence of LLMs not being genuine intelligence[1]. Moreover, the prospect of large-scale high-tech plagiarism driven by LLM technology became a concern that eventually led, last June, to EU regulation 2024/1689 laying down harmonised rules on the use of AI.

Early in December 2022 emails began circulating, more or less surprised at what LLMs could do in computer programming. Tempted to challenge ChatGPT with a problem not widespread in the literature, the author chose a simple problem description from the introductory classes of one of his courses, which goes like this: *"For each list of calls stored in a mobile phone (eg., numbers dialled, SMS messages, lost calls), the store operation should work in a way such that (a) the more recently a call is made the more accessible it is; (b) no number appears twice in a list; (c) only the most recent 10 entries in each list are stored."*

ChatGPT promptly generated a Python program that, despite minor syntax errors, was quite acceptable and not much different from what a first-year student would write. Interestingly, it suffered from unnecessary implementation detail such as recording the exact time of each call — precisely the same kind of mistake that students tend to make in their (very often "biased") programs.

As the course deals precisely with how to build elegant and generic programs using functional programming, the next experiment was to ask for a program for the same challenge, but in Haskell, the language used in the course. Although Haskell is less popular than Python, the software produced a syntactically correct Haskell program capable of using existing libraries, albeit in a rather baroque way.

Impressed with the results, the author responded with an unusual request: as the course addresses a concise programming style that dispenses with program variables to ensure correctness by construction, a solution in this style (termed "point-free") was now requested. The result was quite surprising (see the figure below); quite surprising - because, in the challenge for which it was perhaps least trained, the machine came up with the best solution: a solution with a three-step pipeline, one per paragraph of the problem, in the correct order, with just one error in one of the steps.

Realising how this error contained a technical subtlety that was studied in the course, it was decided to use the example in an exam question[2]: students were asked to analyse the solution proposed by ChatGPT and diagnose the error. The results were anything but encouraging, as this was the question most students sank on, and for a simple reason: they are trained neither to analyse code written by others nor to express themselves in prose. Worse than that, faced with such an unusual question, most of them didn't even try to answer.

1 https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html
2 https://haslab.github.io/CP/2223/Material/cp22231.pdf

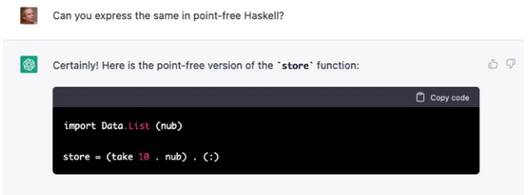
Wes Cockx & Google DeepMind / Better Images of AI / AI large language models / CC-BY 4.0

In retrospect, such a reaction is nothing more than part of a much larger problem in social behaviour, and one that negatively impacts education: the lack of critical thinking caused by the current overdose of social network communication based on telegram-like messages that are forwarded without critical evaluation. This is the crucible in which fake news ferments.

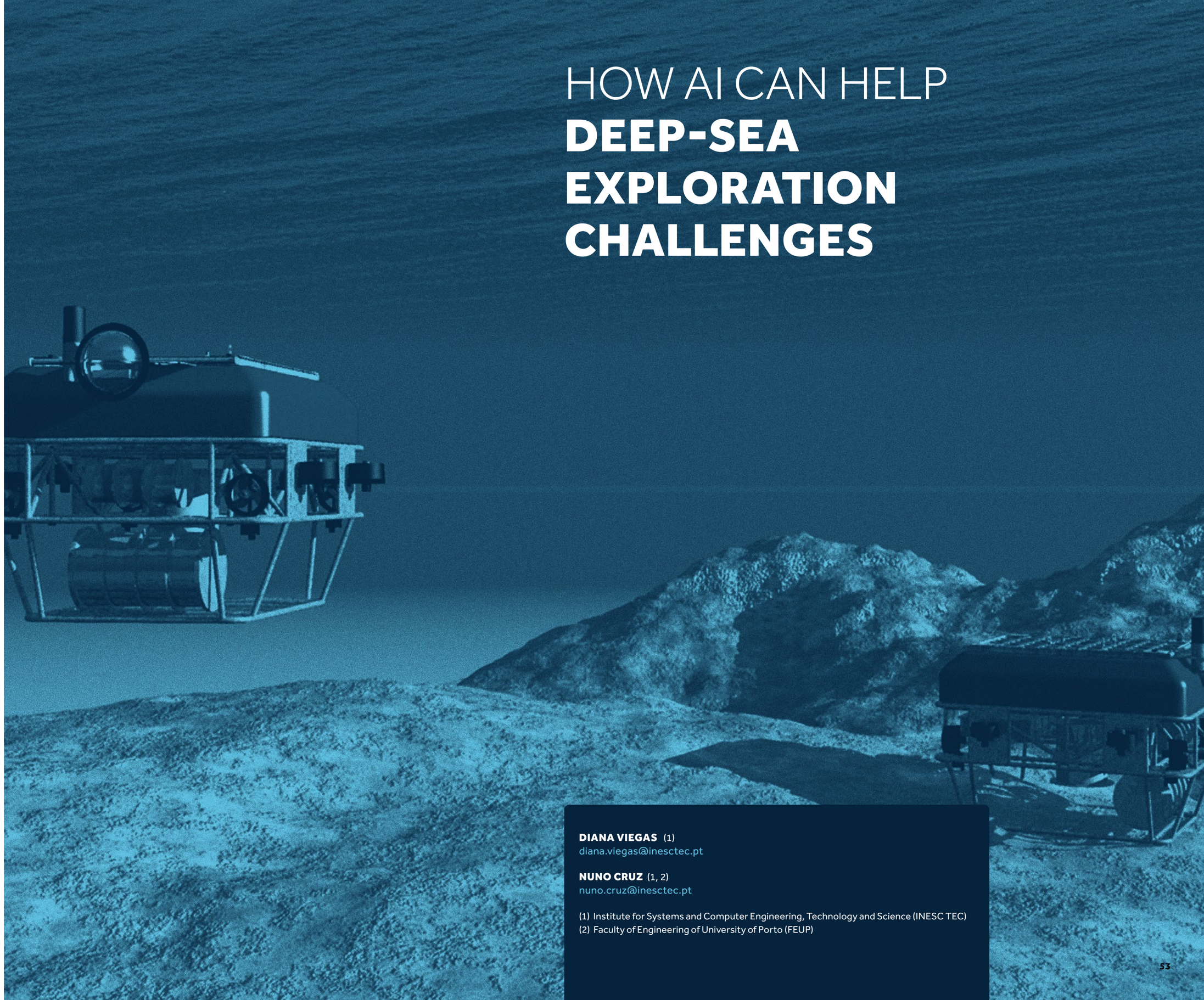
Aware of the urgent need to prepare computer science students for a critical use of LLM-based tools, the author has since embarked on a teaching style in which, upon formulating a new problem, the outcome of a LLM solution (typically by ChatGPT) is promptly taken as starting point and analysed by students. Experience in doing this at master's degree level, for Alloy models generated from problem requirements, has been quite pedagogical. It is true that much garbage is generated, but students are learning from it and becoming aware that better prose generates better models.

It must be said that these experiments have been carried out tentatively, not systematically. Nevertheless, what has been learnt can already be framed in a more subtle setting, a kind of "revenge of the Arts". Why revenge and in what sense? Many students will have sought STEM training to free themselves from unpopular subjects such as literature, poetry, and the arts in general. As a result, they lack command of written prose, let alone articulated speech. If LLMs now seem to need well-written requirements to produce less waste, how ready are students to properly formulate their quests? Is having "survived" reading a voluminous novel like Tolstoy's *Anna Karenina* actually an asset for being a good programmer in the LLM era?? Here is a provocative question deserving some meditation. We live in the offspring of the "big divide" between art and science, in the name of specialisation and productivity. Moreover, the proclivity of youngsters to technology is well-known. But we may have to revisit such a disastrous split in our modern times, as some are already proposing by advocating STEAM and not just STEM education, where the 'A' stands of course for 'art'.

These are odd times for anyone who, like the author, still regards programming as a calculational exercise out of which a correct computer program should emerge. Clarity and economy of thought are essential to such exercises, as coincidentally LLM-generated programming seems to require too. We should not rule out any technology enabling us to produce good software, be it through AI, mathematics or both. One thing can be taken for granted: who in the future will fly in a plane whose LLM-generated software has not been verified 100% correct? On the day direct generation of programs from requirements proves to be definitively effective, the need for (formal) verification will remain. And this is where jobs in the future of computing are likely to be found.



HOW AI CAN HELP **DEEP-SEA EXPLORATION CHALLENGES**



Robotic operations support at sea bottom

DIANA VIEGAS (1)
diana.viegas@inesctec.pt

NUNO CRUZ (1, 2)
nuno.cruz@inesctec.pt

(1) Institute for Systems and Computer Engineering, Technology and Science (INESC TEC)
(2) Faculty of Engineering of University of Porto (FEUP)

THE IMPORTANCE OF DEEP-SEA EXPLORATION

A very significant part of the deep sea is still unmapped, particularly in areas that are difficult to access. Earlier this year, new previously unknown “underwater mountains” were discovered in Chilean waters, teeming with deep-sea life. Exploring the deep sea reveals new species and ecosystems unknown to science, and this expedition brought to our knowledge about 100 new species. These discoveries highlight the potential for finding many more unique organisms, providing insights into how life thrives in extreme environments, which can lead to advancements in biology and ecology, and even provide clues about life on other planets.

Until now, the resolution of available maps was so coarse that these mountains were “hidden.” Besides the morphology, scientists find dozens of new species on each expedition. These discoveries raise many questions that remain unanswered: how do these species live and interact with each other or with their surrounding environment? How have these species evolved in these extreme environments, in the complete absence of sunlight and under immense pressure?

The deep sea holds geological information crucial for understanding Earth’s history and processes. Discoveries like underwater mountain chains provide significant data on tectonic activity, volcanic processes, and crust formation. Mapping these landscapes helps scientists predict natural disasters, such as earthquakes and tsunamis, and develop strategies to mitigate their impact. Moreover, it is undoubtedly an unknown source of raw materials as the green transition is pushing the limits of inland mines for new and high concentrations of specific minerals.

The unique conditions of the deep sea lead to organisms with remarkable adaptations. Studying these can result in breakthroughs in medicine and technology, such as enzymes functioning under extreme conditions for industrial or pharmaceutical use. The deep sea is also a source of novel compounds and materials, driving innovation and contributing to various fields.

Deep-sea exploration is essential for environmental conservation: understanding the biodiversity and dynamics of these areas helps protect them from human activities like deep-sea mining and fishing. Detailed knowledge allows for better conservation strategies, vital for maintaining ocean health, supporting global biodiversity, and ensuring the well-being of human populations relying on marine resources.

But how can we better understand and explore the deep sea?

Deep Sea is posing huge challenges due to the high pressure, the harshness of the ocean, the lack of light and the impossibility of humans going there in safe conditions. We must improve our robotics and autonomous systems to go deeper and stay longer underwater.

The present technological developments available for deep-sea deployment, in addition to the lack of reliability and robustness, are limited by the depth requirement, the level of autonomy, the computer power processing, the endurance, and the lack of reliable underwater communications capacity.

It is important to tackle the technological limitations and gaps and establish advanced methodologies for effective exploration and real-time monitoring of environmental impacts.

HOW CAN AI HELP DEAL WITH THE MAIN CHALLENGES IN DEEP-SEA EXPLORATION?

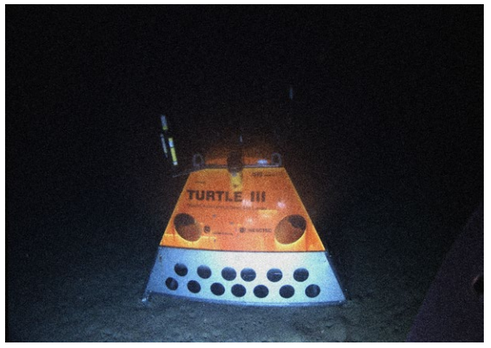


Figure 1
Robotic Lander TURTLE in a mission in the deep sea

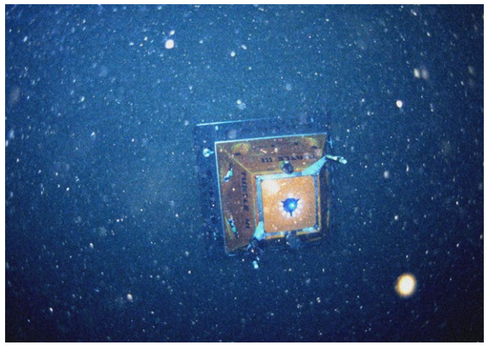


Figure 2
Optical communication link from TURTLE to an autonomous underwater vehicle (AUV) enabling real-time monitoring

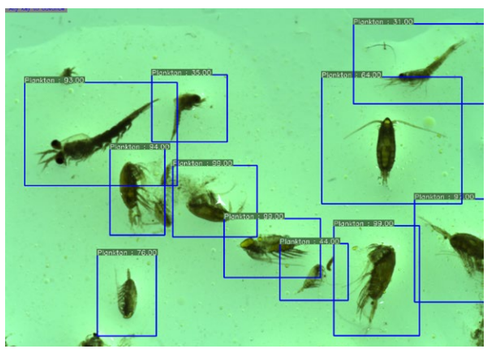


Figure 3
Automatic detection and classification of zooplankton

1. Operating in Extreme Conditions

The deep sea is characterised by immense pressures that increase with depth, requiring specialised equipment capable of withstanding such conditions without being crushed. In addition, near-freezing temperatures also affect electronic equipment, requiring materials and systems designed to operate reliably in cold environments. These harsh conditions increase the likelihood of technical failures; therefore, redundancy, rigorous testing, and robust design are essential to mitigate risks.

By analysing historical data and failure patterns, AI can recommend optimal maintenance schedules. AI tools can analyse data from sensors embedded in equipment to predict potential failures before they occur, reducing the risk of catastrophic failures. For instance, it can monitor pressure sensors, temperature gauges, motion patterns, and power consumption, to identify subtle signs of wear or malfunction.

Sunlight does not penetrate the deep ocean, so video systems require advanced lighting systems and proximity to get meaningful data. AI can process and analyse images and video footage in real-time. Techniques such as machine learning and computer vision can enhance visibility in low-light conditions, identify key features, and even classify objects or species.

2. Technical Operational Constraints

Maintaining communication with underwater equipment is difficult due to the limitations of radio waves and reliance on acoustic signals, causing delays and reduced control capabilities. A simple message takes 10 seconds to receive a reply from a device that is 7.5km away using acoustic signals. Typical autonomous underwater vehicles are programmed to follow pre-determined routes, usually defined as a sequence of waypoints, and all possible scenarios, like facing hypothetical obstacles, must be considered. AI tools can make real-time adjustments to the vehicle’s course and operational parameters based on sensor data, ensuring the vehicle can handle unexpected changes in the environment. Furthermore, AI tools can optimise navigation paths for underwater vehicles, avoiding obstacles and adapting to changing conditions, including the state of performance of thrusters or other onboard equipment. This helps in optimising the exploration of uncharted areas with minimal human intervention.

Another technical challenge is to obtain accurate location of underwater devices and provide an accurate estimate of related measurements. Main techniques rely on underwater acoustics, prone to their own sources of errors, e.g., due to the varying conditions of sound propagation. AI tools can estimate a model of the acoustic propagation channel, based on local measurements, and reduce the location errors.

Finally, all existing autonomous robotic systems, relying on batteries, have limited operational durations and spend a significant part of their energy in descending/ascending to/from the areas of interest. To this end, AI tools can optimise power consumption by adjusting operational parameters (e.g., velocity, active sensors, thruster allocation) based on real-time performance data.

3. Mission Planning, Logistical and Financial Challenges

Exploration activities can disturb fragile ecosystems such as most deep-water environments; therefore, developing sustainable exploration practices and careful planning of deep-sea operations is crucial to ensure that these activities do not harm unique and delicate deep-sea environments. Reaching most deep-sea areas requires significant logistical planning and resources, including launching and recovering exploration vehicles from ships in remote ocean regions. Hence, deep-sea exploration is expensive, involving high-tech equipment, specialised vessels, and extensive support teams. Funding such missions often requires substantial investment from governments, private companies, or research institutions.

AI tools can optimise mission planning, not only by designing routes for complementary robots but also by selecting ideal mission parameters (e.g., velocities and depth profiles) for specific objectives. Moreover, AI tools can be used to analyse global data, find unexpected correlation patterns, and help identify preferred locations for specific goals. With the new developments of AI tools based on large language models, it is possible to anticipate new capabilities for autonomous vehicles where the operator only sets the main objective or research questions, and the robot is fully “driven” by a virtual operator.

4. Data Management

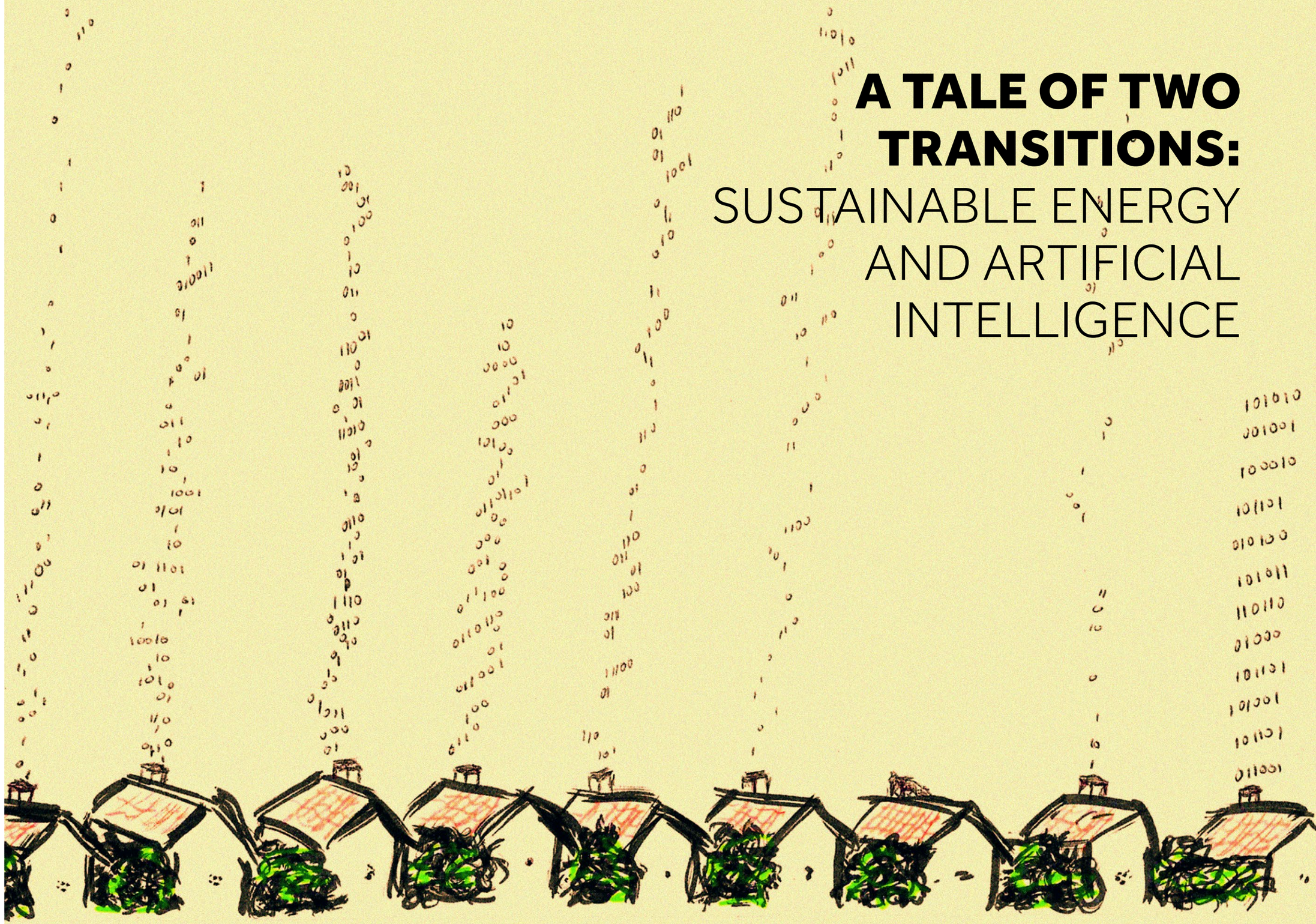
Managing the vast amounts of data collected, including video, images, and sensor readings, is complex and resource-intensive. Effective storage, processing, and analysis are essential for yielding meaningful insights. AI can identify unusual patterns or changes in environmental conditions that could indicate potential issues, such as sudden temperature drops or pressure anomalies, and isolate particular sets of data relevant to a specific study. In other cases, for example in deep-sea observatories, AI can analyse the behaviour of marine life and environmental changes over time, providing valuable insights into ecosystem health and dynamics.

5. Deep water manipulation

Remote-operated vehicles (ROVs) typically operate with a tether that provides immediate situation awareness (mostly with image and sonar feedback), facilitating underwater manipulation. In deep waters, where the cable is unwanted, autonomous manipulation is rarely used, in a very limited scope. AI-driven robotic arms can perform maintenance and repair tasks on underwater equipment. They can operate with precision and adapt to complex scenarios, potentially fixing issues without needing human intervention.

The challenges of Deep-sea Exploration are huge, as well as the challenges for AI tools with a great potential to revolutionise this field. By leveraging AI, we can enhance efficiency, safety, and the overall understanding of deep-sea environments. Advanced data processing, autonomous operations, predictive modelling, real-time monitoring, and improved decision-making, are some of the possible tools that will improve the capacity to increase the knowledge and explore the Ocean, enhancing the efficiency and safety of deep-sea activities also contributing to the sustainable management of ocean resources.

A TALE OF TWO TRANSITIONS:
SUSTAINABLE ENERGY AND ARTIFICIAL INTELLIGENCE



Joahna Kuiper / Better Images of AI / Little data houses / CC-BY 4.0

RICARDO BESSA
Institute for Systems and Computer Engineering, Technology and Science (INESC TEC)
ricardo.j.bessa@inesctec.pt

ENERGY SECTOR LANDSCAPE

Over the last two decades, the energy sector has undertaken a structural transformation summarised by the 3Ds: decarbonisation, decentralisation, and digitalisation.

The drive towards decarbonisation has seen notable progress through increased integration of renewable energy sources. This involves strategic actions, such as replacing carbon-intensive technologies like coal power plants with large-scale renewable energy power plants, increasing renewable energy self-consumption rates among industrial, domestic, and buildings, and electrifying vehicle fleets. Additionally, efforts extend to energy vectors like green hydrogen and energy storage technologies, providing enhanced system flexibility, including seasonal storage, and (at least) keeping the security of energy supply. However, the substantial increase in renewable energy introduces significant challenges in all energy system elements: generation, transmission, distribution, and consumers.

Decentralisation takes place through various actions. This includes distributed generation technologies like co-generation power plants, collective photovoltaic installations, and cold/heat waste reuse, offering local consumers and communities energy at a cost below retail prices. The emergence of the prosumer, a citizen capable of producing and consuming electrical energy, further contributes to decentralisation. Prosumers can buy and sell electricity to the primary grid individually or as part of a local energy community. The evolution of new business models focusing on shared asset ownership or renting requires robust financial mechanisms and regulatory frameworks to ensure energy equity and resilience, especially for vulnerable consumers.

Digitalisation was initially driven by deploying the smart metering infrastructure. However, recent advancements in Internet-of-Things and cloud technology are expanding digitalisation beyond the electrical infrastructure to encompass grid users and service providers, including those from related sectors like mobility. Concepts like digital twins, energy data spaces, and the internet-of-energy are emerging, with several pilot projects currently in progress, meaning a shift towards a more connected and intelligent energy landscape.

APPLICATION OF ARTIFICIAL INTELLIGENCE (AI) IN THE ENERGY SECTOR

The European Commission (EC) white paper on "Artificial Intelligence: a European approach to excellence and trust" describes how a regulatory framework for AI in the European Union (EU) could be developed and classifies the energy sector (among others like healthcare and transport) as a high-risk sector. Hence, this sector has been using expert systems as the core AI technology due to a) their structured and organised way of representing and storing expert knowledge, b) consistent decision-making, i.e., by applying the same rules and knowledge to similar situations, and c) the possibility documenting and transferring expert knowledge. One of the first state-of-the-art reviews was published in 1989, framing AI under the name "expert systems" [1]. Nowadays, expert systems are still available in commercial products and grid automation, e.g., grid protection systems and restoration.

The demand for adaptable solutions that can learn from data — whether collected from field sources or using traditional physics-based software tools for energy system simulation — has significantly increased with the expansion of power systems and the integration of new energy sources. This motivated research in artificial neural networks and other machine learning methodologies, including decision trees and fuzzy inference systems. Initially focusing on power system operation, this research gained momentum as the 21st century began, broadening its scope to encompass emerging applications such as demand response, renewable energy forecasting, battery storage optimisation, and asset management [2]. Examples of cases of success in industry are the use of decision trees and neural networks for dynamic security assessment in Hydro-Québec and BC Hydro power systems [3]; the use of several machine learning models (e.g., neural networks, gradient boosting trees) for short-term wind and solar energy forecasting [4]; predict the distribution network faults that are likely to occur under the given circumstances and their respective repair durations [5]; or, a data-driven system that provides personalised energy efficiency recommendations for commercial customers [6].

Recent breakthroughs in AI research have led to a reinforced use of this technology within the energy sector, such as increased performance and decreasing costs of hardware, advances in deep learning for different areas like computer vision or natural language processing, new paradigms such as transfer learning and generative AI, automated and low-code AI platforms, and brain-inspired AI concepts (e.g., attention mechanism). Moreover, industry-driven challenges, exemplified by L2RPN (Learning to Run a Power Network) from RTE, have prompted collaboration among AI scientists and power system experts [7]. These collaborative efforts motivated different groups to develop a new reinforcement learning-based assistant to aid human operators in operating electrical grids during normal operations and when the system is under stress due to overloads or disturbances.

Two other emerging paradigms in the energy sector are physics-informed machine learning and edge intelligence. In problems where numerical analysis approaches are complex to design or too expensive to compute accurately, machine learning techniques are used to solve algebraic equations or directly handle scenarios with limited data [8]. The need to control locally distributed energy resources or microgrids, or concerns with energy-intensive computing and data security, motivates the research in edge AI for energy systems [9].

To conclude, different energy sector stakeholders are putting their attention to AI technology, namely electricity system operators, energy retailers, energy services companies, consumers/prosumers, communities, software and automation vendors, among others, with the following main drivers for adoption:

- The ongoing structural transitions of electricity systems to accommodate many diversified and distributed energy resources, such as renewables, energy storage, and electric vehicles. For instance, addressing challenges like renewable energy variability and forecast uncertainty demands the creation of innovative tools for operating an energy system. This includes refining load and renewable energy forecasting methodologies and creating novel tools to enhance human real-time decision-making processes.

Image generated with AI using OpenAI's DALL·E

- The evolution of electricity markets with increasing market actors and services diversification. Planning under these changes can be facilitated through new digital technologies. For instance, AI helps achieve decision-making automation in emerging local energy communities, e.g., peer-to-peer energy trading.
- New challenges to system resilience, e.g., considering climate change and man-made hazards like cyber-attacks, could be mitigated through the integration of different data sources and the use of digital technologies. For instance, AI can augment policymakers' analytical capabilities, e.g., derive interpretable rules to explain energy scarcity events [10].
- Increasing potential to analyse and optimise electricity demand patterns on the consumer side, e.g., through smart meters, controllable devices, and building sensors. AI can create socially relevant products, such as energy poverty forecasting or energy efficiency recommendation systems.

A JOURNEY TOWARDS AN INTERDISCIPLINARY RESEARCH AND INNOVATION ECOSYSTEM

The CIGRE Working Group C2.42 has established an innovation roadmap that guides the research community toward goal-oriented advancements in AI. This roadmap aims to leverage AI's potential while ensuring high-quality testing and safety standards. The strategy includes three main components: a) fundamental research to establish proof of principles, b) open-source initiatives for proof of concepts, and c) testing and experimentation facilities (TEF) for the integration and industrialization phases.

According to the EC definition, a TEF is a "combination of physical and virtual facilities, in which technology providers can get primarily technical support to test their latest AI-based software and hardware technologies (including AI-powered robotics) in real-world environments."

INESC TEC's work in AI for energy systems, as depicted in Figure 1, aligns with this roadmap and the EU AI strategy and AI Act.

Figure 1 INESC TEC's ecosystem in AI applied to energy systems

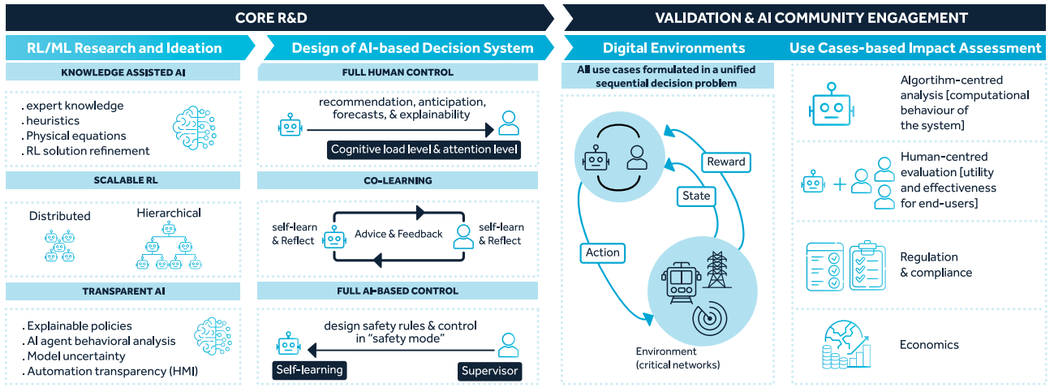


Figure 2 AI4REALNET research approach

In fundamental research, INESC TEC leads the Horizon Europe AI4REALNET project (see Figure 2 for the project's research approach), which applies AI to critical infrastructures such as power grids, railways, and air traffic management. The project aims to improve human decision-making with AI support rather than simply deploying AI systems. The goal is to optimise the collaboration between humans and AI, enhancing the overall efficiency of socio-technical systems and ensuring consistent human engagement and performance. This interdisciplinary approach involves traditionally separate fields, such as philosophy, psychology, and human reliability, to study how experts make collaborative decisions in complex situations and develop effective design and evaluation criteria for supporting human decision-making. It also includes cognitive and biomedical engineering to understand human cognitive processes and improve human-machine interfaces, as well as physics, mathematics, decision theory, computer science, and specific engineering domains related to energy and mobility.

The ENFIELD, a European AI Networks of Excellence Centres, integrates different disciplines, including Green AI, Adaptive AI, Human-centric AI, and Trustworthy AI. Here, INESC TEC is advancing human-centred AI research to develop inherently interpretable AI models. These models are designed to be transparent, allowing humans to understand and adjust the mechanisms that convert inputs to outputs, particularly when system behaviour deviates from expectations. Current research focuses on developing evolving expert systems capable of learning and improving from data and tackling tasks such as supervised and reinforcement learning, e.g., classifying the dynamic security of power systems or designing optimal control strategies.

In the Horizon Europe Green.Dat.AI project, the energy consumption of AI methods is at centre stage, and INESC TEC is developing federated learning and edge AI techniques for smart electric vehicle charging and renewable energy optimisation, as well as a methodology and software for monitoring the energy consumption of AI-based methods.

This research is supported by an by an open source initiative across INESC TEC, promoting innovation and collaboration by making algorithms available to the broader community. This initiative contributes to the AI-on-demand platform, accelerating AI advancements and fostering transparency.

In the industrialisation phase, INESC TEC uses two key instruments: TEF and the Digital Innovation Hub (DIH). In October 2024, INESC TEC started to establish nodes in two European TEFs for local energy communities/microgrids (AI-EFFECT) and marine renewable energy (enerTEF). These nodes will support integrating, testing, and demonstrating cutting-edge AI technologies in the energy sector in collaboration with local partners like *Cooperativa Eléctrica do Vale d'Este and Companhia da Energia Oceânica*.

To support start-ups and SMEs in enhancing their products, services, and processes using digital technologies like AI and high-performance computing, INESC TEC coordinates a DIH called ATTRACT. This hub provides technical expertise and domain knowledge across various sectors, including energy and infrastructure, and provides innovation services to aid industry transformation. In the energy sector, it helps design and test "quick-win" AI use cases and validates technology readiness levels, leveraging the capabilities of the TEFs.

Finally, INESC TEC's involvement in European associations like ADRA and AIOTI enables the institution to contribute to European AI and innovation policy while refining in-house internal research and innovation objectives. This engagement aligns INESC TEC's projects with broader EU priorities, facilitates collaboration, and provides access to AI's latest developments and funding opportunities.

CONCLUDING REMARKS

Modern AI technology can bring value to the energy sector across different dimensions. Firstly, fast decision-making in operating and planning power systems with high shares of renewables, where flexibility from various sources (generation, consumers, or grid assets) is fundamental. This is especially crucial under challenging scenarios, such as extreme weather events and cyber-attacks, where the system's adaptability becomes instrumental in maintaining infrastructure/system integrity and resilience. Secondly, it will enable the optimal operation of new business models, such as energy sharing between prosumers, smart electric vehicle charging, and de-risk investment in energy efficiency actions. This will contribute to democratising access to renewables at an affordable cost. Thirdly, it can systematically process, explore, and exploit large volumes of heterogeneous data spanning the entire energy value chain and beyond, encompassing mobility, water, and high-performance computing domains. Thus, it will enhance and potentially automate existing (or new) tasks and processes traditionally handled by humans or expert systems with new requirements like adaptability and robustness to new scenarios.

Nonetheless, the energy consumption associated with AI solutions demanding extensive computing resources is a significant concern for two sectors — energy and high-performance computing — both actively advocating for complete decarbonisation and rational electricity use. Notably, the industrial deployment of large language models requires substantial computational resources, leading to increased energy consumption. Data privacy and security are also primary requirements for AI since, in various use cases, personal data (e.g., energy consumption, in-door sensors, outage events) or confidential data about the network infrastructure or electricity market trading are used. Therefore, it is necessary to create robust solutions to data breaches where the reliability and security of the AI model are paramount. Certification and formal verification of AI models that operate autonomously or provide recommendations to humans is essential to guarantee trust.

Image generated with AI using OpenAI's DALL-E

ACKNOWLEDGMENTS

This work was supported through projects AI4REALNET (GA No. 101119527), ENFIELD (GA No. 101120657), AI-EFFECT project (GA No. 101172952), and enerTEF (GA No. 101172887); all funded under European Union's Horizon Europe Research and Innovation programme. Views and opinions expressed are, however, those of the author only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible. The author acknowledges all members of the CIGRE C2.42 working group for discussions.

REFERENCES

1. Zhang, Z. Z., Hope, G. S., Malik, O. P. (1989). Expert systems in electric power systems – a bibliographical survey. *IEEE Transactions on Power Systems*, 4(4), 1355-1362.
2. Kezunovic, M., Pinson, P., Obradovic, Z., Grijalva, S., Hong, T., Bessa, R.J. (2020). Big data analytics for future electricity grids. *Electric Power Systems Research*, 189, 106788.
3. Huang, J. A., Valette, A., Beaudoin, M., Morison, K., Moshref, A., Provencher, M., Sun, J. (2002). An intelligent system for advanced dynamic security assessment. In Proceedings. *International Conference on Power System Technology* (Vol. 1, pp. 220-224). IEEE.
4. Bessa, R. J., Möhrlen, C., Fundel, V., Siefert, M., Browell, J., Haglund El Gaidi, S., et al. (2017). Towards improved understanding of the applicability of uncertainty forecasts in the electric power industry. *Energies*, 10(9), 1402.
5. Vähäkuopus, S., Paananen, H., Anttila, L., Kupila, T. (2019). Predicting the impacts of the major disturbances for better resource management and situational awareness. *25th International Conference on Electricity Distribution (CIRED 2019).*
6. Zawadzki, P., Lin, Y., Dahlquist, F., Bao, T., Laurain, A-L., Johnson, K. (2016). Personalized energy efficiency program targeting with association rule mining. *ACEEE Summer Study on Energy Efficiency in Buildings*.
7. Marot, A., Donnot, B., Dulac-Arnold, G., Kelly, A., O'Sullivan, A., Viebahn, J., et al. (2021). Learning to run a power network challenge: a retrospective analysis. *NeurIPS 2020 Competition and Demonstration Track* (pp. 112-132). PMLR.
8. Stiasny, J., Chatzivasileiadis, S. (2023). Physics-informed neural networks for time-domain simulations: Accuracy, computational cost, and flexibility. *Electric Power Systems Research*, 224, 109748.
9. Himeur, Y., Sayed, A., Alsalemi, A., Bensaali, F., Amira, A. (2023). Edge AI for internet of energy: challenges and perspectives. *Internet of Things*, 101035.
10. Heymann, F., Bessa, R.J., Liebensteiner, M., Parginos, K., Hinojar, J. C. M., Duenas, P. (2022). Scarcity events analysis in adequacy studies using CN2 rule mining. *Energy and AI*, 8, 100154.
11. Cremer, J. L., Kelly, A., Bessa, R. J., Subasic, M., Papadopoulos, P. N., Young, S., Marot, A. (2024). A pioneering roadmap for ML-driven algorithmic advancements in electrical networks. *IEEE PES ISGT Europe 2024*

AI AND SUSTAINABILITY: THE OPPORTUNITIES AND THE RISKS WE FACE

Photo credits: Google DeepMind | Pexels

ANTÓNIO BAPTISTA (1)
antonio.baptista@inesctec.pt

ANTÓNIO LUCAS SOARES (1, 2)
antonio.l.soares@inesctec.pt

(1) Institute for Systems and Computer Engineering, Technology and Science (INESC TEC)
(2) Faculty of Engineering of University of Porto (FEUP)

The tools and concept associated with "Artificial Intelligence" have gained significant prominence among consumers over recent years, as well as great business interest among investors and companies. However, and like any "El Dorado", the implications of these new solutions must be analysed and safeguarded - whether in terms of economic, environmental and social sustainability.

The dawn of significant advances in the transposition and improvement of mathematical algorithms framed within the broad concept of "Artificial Intelligence" (AI), namely with the current high computing capacities and large amounts of data for model training, allowed the emergence of multiple AI-based applications to solve complex problems (e.g., health, energy, industry, transportation, marketing, etc.). The success of those applications and developments led to two global effects: the interest of markets and business investors, and the wide dissemination of this new paradigm to consumers and citizens (namely through Large Language Models, popularised by the ChatGPT solution, by the company OpenAI).

The large investments of the global giants of the software industry are clear and we can say the same about hardware manufacturers (chips, large computing and storage systems). Still relevant, but significantly unbeknownst to the consumers, is the need to intensive consumption of electricity to cover computing power and related cooling systems' needs, and to address the water used in the cooling infrastructures. Indeed, in this new "El Dorado" (as well as previous ones) there is a natural enthusiasm regarding the discovery and the search for "new mines" of opportunities and applications, towards hopefully the improving of regions and population. Hence, and still concerning the context of AI applications, the reader may ask: "aren't these AI or AI-based apps another type of software capable of automating processes through a dematerialisation supported by digitalisation and the Industry 4.0 paradigm?" The answer is "no"; since there are many pros and cons, and we should analyse this question through "sustainability mindset" approach of triple bottom line: economic, environmental, and social domains. This text does not seek and can't

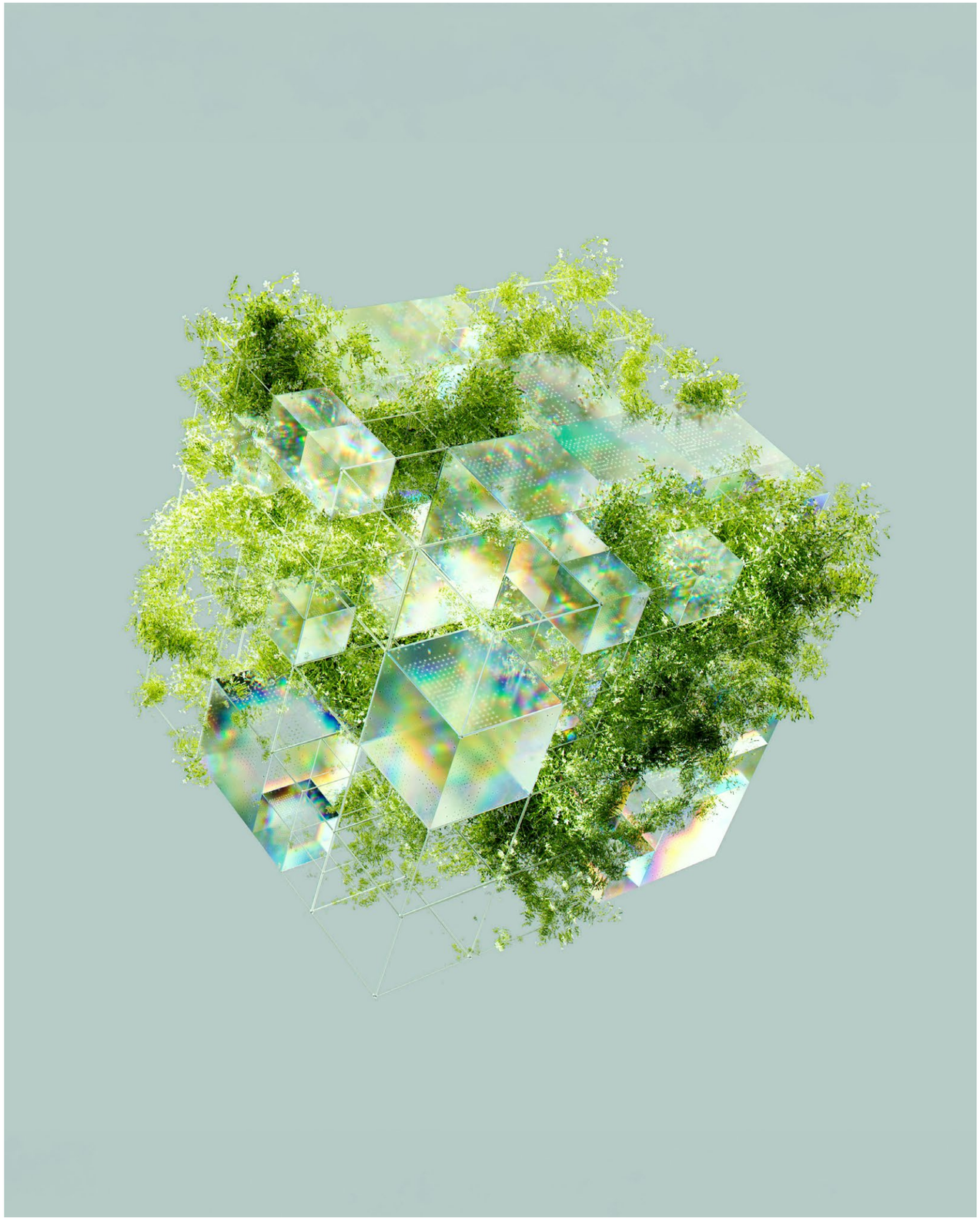


Photo credits: Google DeepMind | Pexels

provide "all the answers" to this analysis. But, it aims to emphasise, through a science-impartial lens, benefits-opportunities and disadvantages-risks, while presenting questions that favour a deeper scientific reflection and improve the citizens' critical awareness.

During each innovation and technological cycle, or even in our history regarding the connection between scientific-technological advances and innovation, there is the usual motivation to seek the capitalisation of results, potential impact in economics and market exploitation terms. Hence, the significant investments made in less than a decade around the new AI paradigm - the "current hype" -, the speculative risks, stock markets and business corrections that will follow (already clear while writing this article) come as no surprise. In this sense, considering economic and financial sustainability, and the broad application of AI solutions - whether in mobility-transportation (e.g., highly anticipated full-autonomous vehicles), industry, or other services -, companies must be careful about their investments and evaluate the potential returns e.g., productivity gains or cost reductions, versus technical (i)maturity, might be for internal processes, products, services or product-services.

In terms of environmental sustainability, natural resources management and pollution levels, it is becoming clear that there are several risks associated with the material, energy and water needs by the applications' operation - despite the benefits of using AI tools and techniques to improve and optimise processes, materials, products, and services (e.g., improved energy or material resources efficiency). We live in a world that currently faces a Triple Planetary Crisis, where the effects of climate change, pollution and biodiversity loss are already clear and tend to threaten the future development of humanity, other species with risk of extinction, and massive destruction of habitats and natural ecosystems. Despite several national or global programmes (notably the Kyoto Protocol 1997 or the UN Agenda Paris 2015) to mitigate the effects of global warming, we regularly witness phenomena like sequential record temperatures, greater frequency, duration and intensity of heat waves, storms and natural disasters. The goal of keeping global warming below +1.5°C may be eminently compromised in view of the limited results of global decarbonisation of the economy [1]. In this unfavourable context, it is increasingly urgent to reduce the emission of greenhouse gases, namely CO2, to accelerate the energy transition to renewable sources, but also to act in terms of water resources management. Concerning energy, the proportionality

Photo credits: Google DeepMind | Pexels

between the computing capacities and the required power, the training of increasingly larger models (with a level of exponential computing requirements), and the mass use of AI tools (namely Generative AI) in professional or personal contexts, translates into higher electricity consumption. As an example, the set of data-centres, computational providers and data-transmission networks represent about 3% of the world's energy consumption, which means an annual CO_2 emission equivalent to Brazil, and almost doubling electricity consumption over four years forecast - from 460TWh, in 2022, to 1000 TWh, in 2026 [2]. The energy level required is so great, that giant software companies are planning or already signing direct energy contracts with nuclear plants, hydro plants or other power plants, while reviewing their decarbonisation plans for potentially slower roadmaps towards carbon neutrality [3]. Considering freshwater consumption, the most recent data are worrying, due to the growing needs related to the cooling of data and computation-centres, but also due to the exponential demand for chips and, consequently, for greater amounts of purified water to feed the manufacturing processes - which compete with the supply of drinking water to the populations. As an example, safeguarding the fact that there are still few scientific studies in the area, recent estimates point to a consumption of 0.5 L of water (associated with the cooling systems of computation centres) for an interaction of 20-50 questions-answer with a Generative AI application [4].

Despite these risks - or disadvantages, if we consider the context and urgent need to mitigate climate change -, the advantages of AI tools cannot be ignored or overshadowed, also in favour of reducing resource consumption (energy, water and materials) in different domains: industry application, city management systems, public transportation, energy distribution networks, fostering and enhancing circular business models, industrial symbioses, waste management, etc. A question that we should consider is whether the application of AI tools and capabilities should focus on important society domains and needs, while informing other agents - namely, citizens - about a responsible and regulated use of AI based tools. In this sense, should the extensive use of AI application be controlled/regulated, or not? Especially since we haven't ensured sustainable energy consumption (with the elimination of non-renewable energy sources) and access to freshwater for all population.

Finally, and regarding social sustainability, there are different types of opportunities, like improving human skills (regular or reduced due to disease/disability), but also risks or threats. This article does not intend to thoroughly explore said threats, since Ethics is a quite sensitive area - addressed in other articles included in the Science & Society magazine; but we must mention the risks associated with a future where people, as intelligent individuals, have access to machines with similar (or higher) cognitive intelligence and decision-making levels, with the power to kill or control communities. When it comes to "less futuristic" aspects - even those related to environmental and economic sustainability - it's worth mentioning the fierce “competition” for resources among the infrastructures that support AI tools, leading to social pressure (namely in the poorest and drought-stricken territories), as well as the consequences in terms of employment due to a broad automation of tasks (many of them still quite specialised today). Thus, similarly to the past, this requires transformation/adaptation processes within organisations, focusing on their professionals and human dimension.

REFERENCES

1. https://www.scientificamerican.com/article/were-approaching-1-5-degrees-c-of-global-warming-but-theres-still-time-to/
2. https://www.forbes.com/sites/arielcohen/2024/05/23/ai-is-pushing-the-world-towards-an-energy-crisis/
3. https://www.ft.com/content/61bd45d9-2c0f-479a-8b24-605d5e72f1ab
4. https://www.euronews.com/green/2023/04/20/chatgpt-drinks-a-bottle-of-fresh-water-for-every-20-to-50-questions-we-ask-study-warns

INESC TEC

INSTITUTO DE ENGENHARIA
DE SISTEMAS E COMPUTADORES,
TECNOLOGIA E CIÊNCIA

Campus da FEUP
Rua Dr. Roberto Frias
4200-465 Porto
Portugal

www.inesctec.pt
science-society.inesctec.pt
info@inesctec.pt
T +351 222 094 000

PRR Plano de Recuperação e Resiliência
REPÚBLICA PORTUGUESA
Financiado pela União Europeia NextGenerationEU
MC MISSÃO INTERFACE
ANI AGÊNCIA NACIONAL DE INOVAÇÃO