More than just AI Skills
Operationalizing trutsworthy AI

Volume 5, No. 6

December 2024

Michael Leyer University of Marburg

Lilian Do Khac University of Marburg

Sarah Engel University of Marburg

Layout & Design: Oliver Behn

White Paper Series of the Chair ABWL:
Digitalisation and Process Management

Volume 5

Introduction

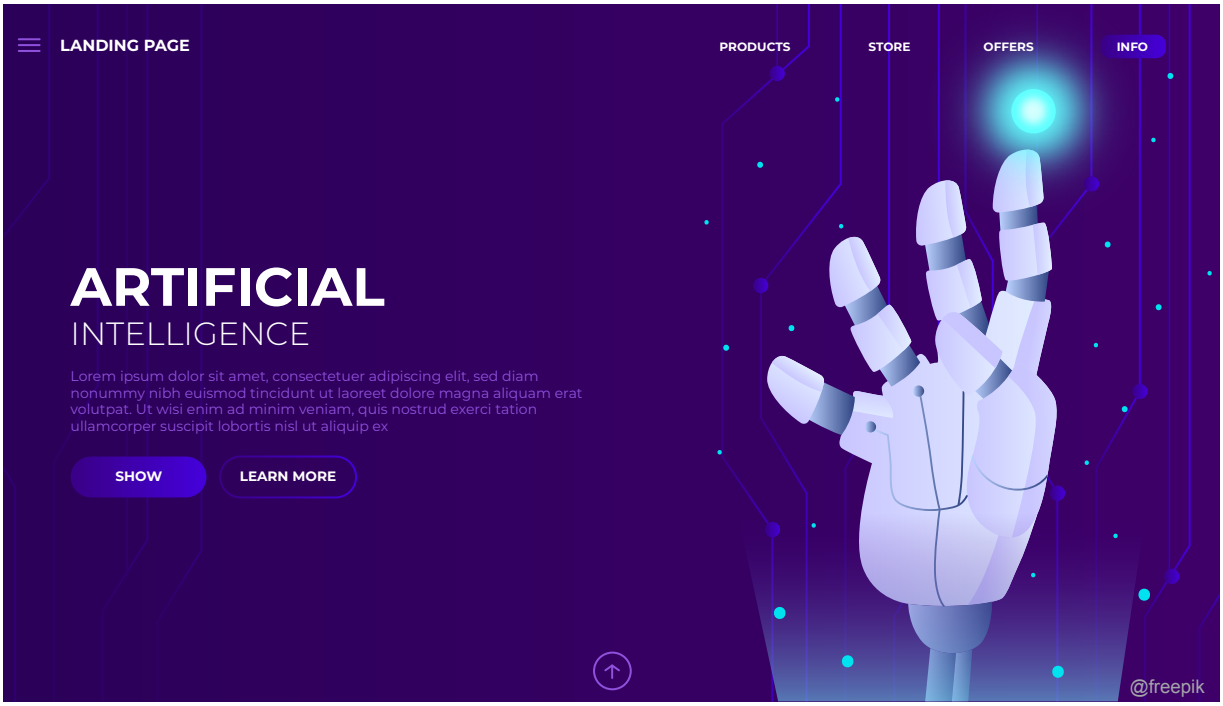
Enforcing quality standards in the development and deployment of AI systems that integrate beneficially into the human and natural fabric will likely be one of the defining questions of the digital age. Already, significant regulatory instruments have been put in place to safeguard fundamental interests such as health, safety, and basic rights, thereby ensuring a functioning social market economy. It is about more than just identifying responsibilities („Responsible“) in a task-focused manner; it is about the commitment and accountability towards the outcome („Accountable“), which is result-focused.

The latter is primarily addressed in the European context under the overarching term „Trustworthy AI.“ AI systems can be deployed in nearly every industry and task, bringing with them vast potential and risks. To manage these and fulfill our duties of care, knowledge of the specific levers directly related to the AI system itself, as well as an appropriate governance framework for such systems, is crucial. A certificate of compliance with quality standards for AI products or services results in trustworthiness and thus expresses competitiveness in the age of digitalization.

A framework for AI-systems

In the relationship between humans and AI, we ask: What qualities must an AI possess to maximize its chances of gaining human trust? In this context, trust is built by the receiving party through the three trust-building factors of abilities, benevolence, and integrity of the sending party. The recipient‘s own attitude also plays a role in determining how successful these factors can be. A review of the academic discourse on AI or autonomous systems in relation to trust suggests that many levers for optimizing (semi-) autonomous systems are already well-known.

From the human perspective, conceptual elements such as robustness, reliability, or multimodality contribute to a positive perception of an AI‘s capabilities. Robustness and reliability have been extensively discussed and are

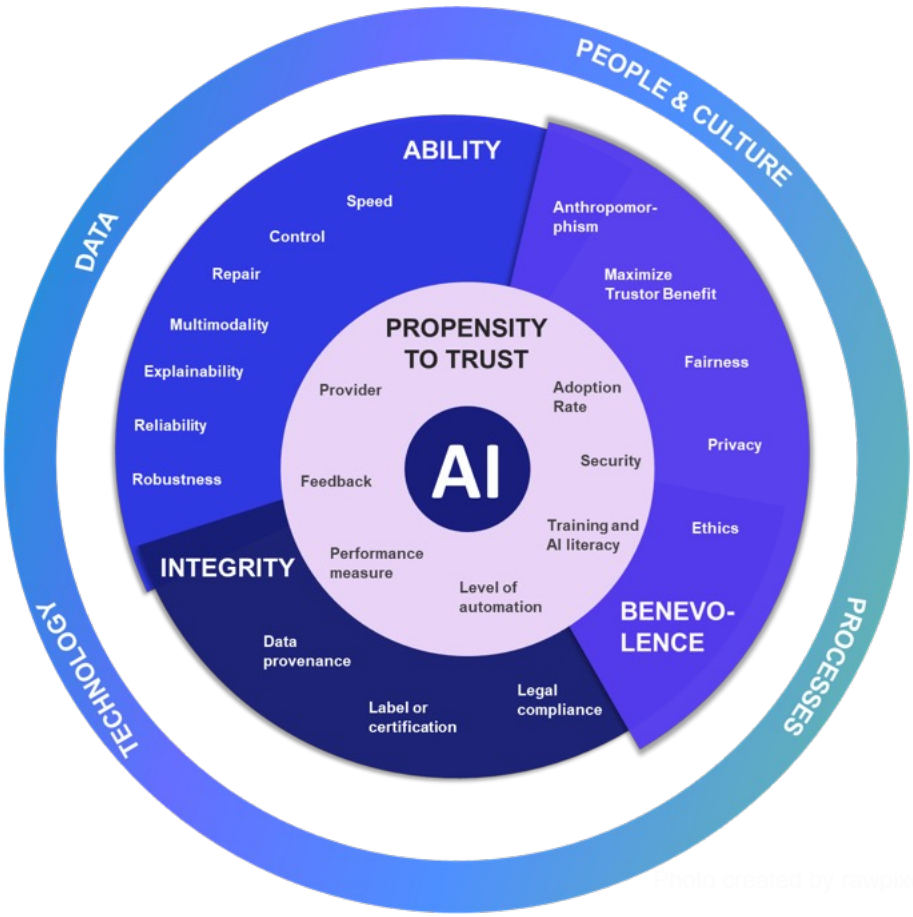


deeply embedded in existing research and literature. Multimodality, on the other hand, is a less explored concept but has shown success with newer AI generations that can process various sensory-motor input channels (text, image, audio). Conceptual elements like anthropomorphism, privacy, or fairness influence a human's positive perception of an AI's benevolence. Anthropomorphic properties refer to human-like form and behavior. Perceived integrity of an AI can be established through data provenance or certifications. Data provenance provides transparency about the training data used for the AI system and serves as a testament to the trustworthiness of this data. Such labeling exists, for example, in the food industry. The internal likelihood of trusting an AI is influenced by conceptual elements such as training and education about the AI systems or the manufacturer behind the AI.

There are numerous levers, but it remains unclear how a suitable calibration set for these levers can be determined for different occasions, domains, etc., or what the fundamental considerations should be.

The right measure for AI governance

The introduction of AI systems has a particularly profound impact on business processes and the interplay of organizations compared to previous technological innovations, as it reinforces the power dynamics between technology and human operators. Unlike humans, AI systems cannot have intentions, so ensuring accountability for their actions requires tracing responsibility across multiple stakeholders, from conception to use. When AI systems are sub-optimally integrated into business processes, it is unlikely that the expected efficiency



gains will materialize, and potential risks may even escalate, undermining the benefits of AI utilization. To ensure that distributedly responsible actors interact seamlessly, transparency and a collaborative framework are necessary to coordinate this complex interplay. This is the traditional challenge that requires a governance structure. Thus, a functioning AI governance framework ensures that trustworthy AI systems are developed and continuously improved. As such, AI governance should guide an organization‘s use of AI and its deployment to achieve its goals. To achieve this, it is essential to be able to assess AI systems for trustworthiness.

The four essential organizational key capabilities for operating a functional AI governance are: data, technology, human and culture, and processes. Data forms the foundation for AI applications. Therefore, mastering data quality and managing data sources is a core competency for successful AI applications. Technology brings an AI application to life. Competence in the right framework of an AI system with the appropriate calibration set is significant. Subsequently, the organization, with its employees and culture, must be able to leverage these AI applications profitably, which can then be scaled through strong processes.

Next steps into the AI age

This framework for trustworthy AI can serve as a guide to navigate your company‘s AI journey and ensure successful implementations. By considering our proposed levers, targeted improvements in human-machine interaction can be achieved. Businesses should be aware of the various elements and their maturity level regarding trustworthy AI. Balancing the control of these activities is crucial to achieving the desired outcomes. Our framework provides comprehensive guidance that enables the building of trust and the development of a positive human-AI relationship. In implementation, companies should consider these aspects to

@freepik

achieve a trustworthy and successful AI integration, leading to a competitive advantage and improved customer satisfaction. Pay attention to control, explainability, reliability, and robustness to establish a solid foundation for your AI strategy, as well as the AI governance framework regarding data, technology, humans, and culture, along with the associated processes.

CONTACT

Prof. Dr. Michael Leyer
Chair ABWL:
Digitalisation and Process Management

School of Business and Economics

Adjunct Professor, School of Management,
Queensland University of Technology,
Brisbane, Australia

Email michael.leyer@wiwi.uni-marburg.de