The *trusted AI* orchestra

Building trust in an AI agent world

Capgemini

Company boards are reshaping corporate strategy to give generative AI freer rein in decision-making. They are considering how AI could be responsible for as much as half of business decisions in the next decade and the fundamental effect that will have on their operating models.

Large language models are now sophisticated enough to be able to beat human opponents in debates through persuasive argument.[1] Gen AI agents have already taken on many business operations, but company management has yet to reach the point of granting AI agents a seat on the board. For the foreseeable future, human intervention will continue to be the highest authority in AI decision-making hierarchies.

The question for business leaders is, where in their organization should they position this human intervention, and where can generative AI be allowed to play its part?

Trust in AI comes in two forms

- *What you trust it to do*
- *How its trustworthiness can be proven*

There are degrees of trust in generative AI's output. It can range from: one where guardrails keep agents compliant but are not so obstructive as to make it ineffective or unviable; to zero tolerance of any hallucinations, as is the case with military intelligence.[2]

Business leaders must decide how to make generative AI risks compatible with their organization's ethical code at the highest governance level rather than view it as a software service operating in isolation.

[1] Salvi, F., On the Conversational Persuasiveness of Large Language Models: A Randomized Controlled Trial, EPFL, March 21, 2024

[2] Economist, The, Babbage podcast, "How spooks are adapting to the digital age", July 10, 2024

Getting the digital corporation ready for *ethical multi-agent Gen AI*

Enterprises have used architecture frameworks such as *TOGAF* for decades. With such a structure, a digital organization operates across four pillars: business, data, applications, and technology. This structure has been effective for AI when governance and planning were by need and per solution with single AI agents. However, with generative AI expanding into numerous functions, there is more risk of conflicts between pillars if they operate multiple agents. This indicates the need to thoroughly consider where generative AI governance sits in organizations.

Trusted AI principles are a superset of an organization's AI compliance and responsible and ethical AI rules. They should be established to anticipate unintended consequences and unexpected risks. Ethics may be the foundation of trusted AI, but trusted AI needs an actionable framework.

Principles of trust in a *multi-agent AI model*

The following fundamental principles should guide how an organization uses generative AI:

01 **Clear identity**
AI use is clearly signposted, including in human interfaces

02 **Cybersecure**
Demonstrable cybersecurity resilience to all probable forms of attack.

03 **Accountability**
There is a clear line to human accountability.

04 **Designed for failure**
An AI agent must be assumed to be capable of failure and have clearly demonstrated failure plans.

05 **Transparency**
Clear and auditable at the agent and system level, with justification and verification for all actions, including the data underlying decisions.

06 **Demonstrably compliant**
An AI agent must be able to positively demonstrate how it is compliant with all relevant laws and regulations

07 **Human control**
AI agents must have explicit human authorization for all interactions with other agents and humans, with opportunity for intervention and overruling at any point

08 **Technically robust**
The ability to work within a contract and defined boundaries, including failure approaches and mitigations.

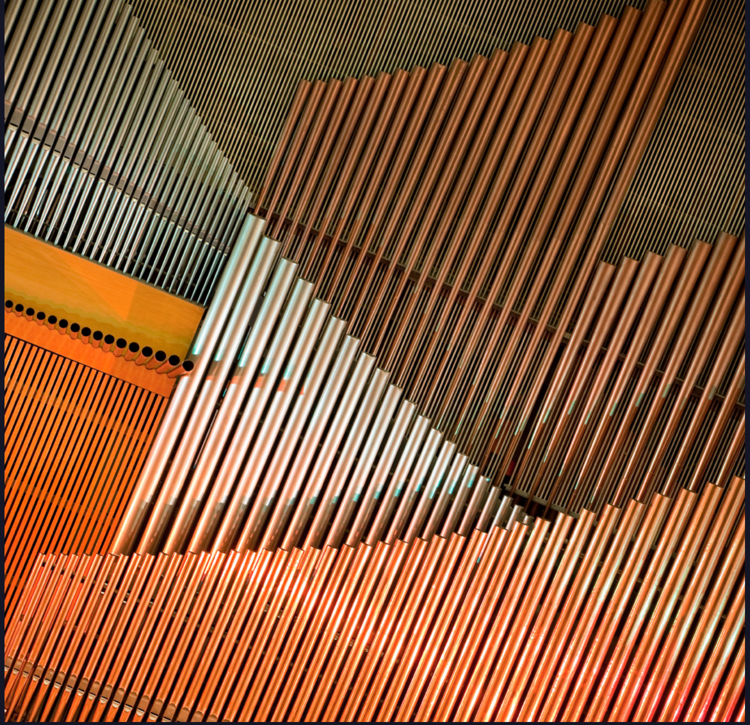
09 **Frugality**
An AI system should complete its operations in the most efficient and sustainable manner. Impact is systematically measurable.

10 **Data quality**
May be managed throughout its entire lifecycle, including feedback mechanisms. This must include data governance controls and testing for bias and fairness

11 **Transparency**
Clear and auditable at the agent and system level, with justification and verification for all actions, including the data underlying decisions.

12 **Tested for failure**
All AI agents must be tested for failure, bias, fairness, as well as all other operational aspects. Testing should be established to identify points of failure and define boundaries.

13 **Privacy/data protection**
Access to and processing of information must be clearly managed, described, auditable, and compliant with privacy regulations. Leakage vectors should be prevented, or mitigations should be demonstrated.



Digital description is *digital clarity*

A company’s digital existence reflects the words of philosopher Ludwig Wittgenstein – “The limit of my language means the limits of my world.” Unless a business model is clearly and comprehensively described digitally, AI will not be able to optimize it. In such a partial state, it cannot be trusted to make decisions since it will not have a complete view of the business nor its constraints, e.g., if a business cannot put into digital form all requirements in a supplier negotiation and all aspects required to judge its success, then an AI agent cannot be trusted to take over a negotiation with certainty of a trustworthy outcome.

Failing to comply *deliberately*

It may seem counterintuitive, but failure is desirable when testing Gen AI. Traditional IT systems fail in conventional ways, which makes testing for failure a matter of identifying the non-success condition.

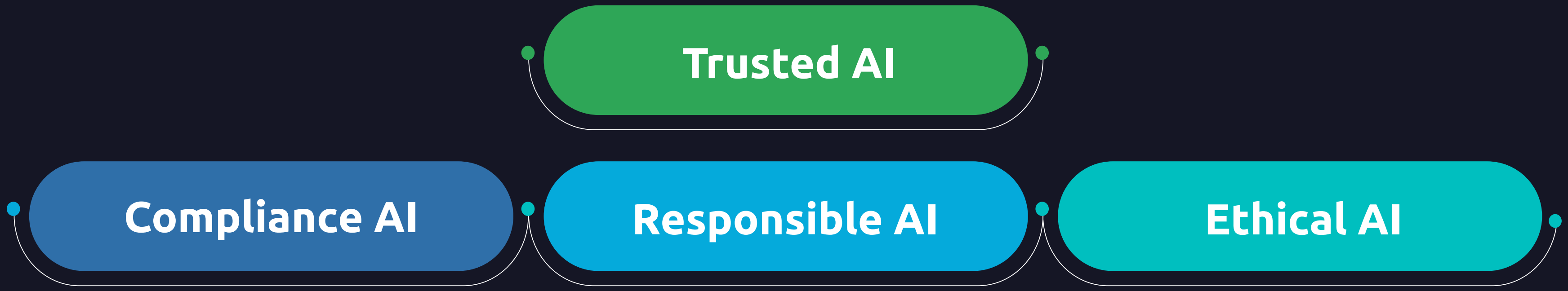
With Gen AI, finding that condition is complicated by hallucinations, prompt injection,[3] model poisoning,[4] and feature expansion. A Gen AI solution might not simply fail to do a task – it could also perform completely unrelated or undesired behavior. As a simple example, most large-scale models boast of their ability to play chess, but no enterprise wants its costly, hyper-capable LLM to expend AI credits through the customer service center because customers have unintentional access to a chess function.

Gen AI testing aims for the system to be demonstrably compliant. To do so, it’s necessary to push to failure, i.e., design for failure and test for bias to establish a model’s boundaries. This reduces the risk of creating false guardrails that could still lead to unexpected breaches or accidents.

[3] A cyber attack used to misdirect a generative AI system.

[4] Tampering with model architecture, training code, hyperparameters, or training data.

Trust must bring together all aspects of AI



Trusted AI

Compliance AI

Responsible AI

Ethical AI

01 A multi-dimensional framework for dimensions, principles, and other aspects

AI must follow applicable regulations and ethics

If Gen AI is to be granted the power to reorganize a call center, create a building management system, or implement corporate strategy, it must have explicit instructions that align with the core principles of the organization.

In delegating decision-making, organizations have to consider the first principal-agent problem, when those creating an AI agent are separate from those commissioning it (the sponsor versus the developer), a problem that is not unique to AI.[5]

There is a following second principal-agent problem, where an AI system is allowed degrees of autonomy, a step further away from human control.

Guardrails control the extent of such a system's model-driven decision-making. At the model level, a business would exclude certain Gen AI systems to prevent potential Trojan horse systems from being allowed into an organization, i.e., an accidental backdoor.

[5] Bostrom, Nick, Superintelligence, Paths, Dangers, Strategies, Oxford University Press, pp. 154-155

02 Trusted AI - business model and strategy

AI must align with organizational strategy, corporate values, and policies to create real value for users and the organization.

At the business model and strategy level, a company sets its principles for guardrails and how accountability will be managed should something go wrong. At its simplest, this means setting a rule, for example, that AI agents will not be permitted to play computer games with clients. That rule would percolate to all current business units and any future acquired entities.

More specific to a corporate context, a company may decide in principle to allow an AI agent to generate and submit Sarbanes-Oxley Section 404 compliance reports, or another jurisdiction's equivalent, on behalf of its corporate officers. Here, the company would also designate who is accountable should a report be rejected by the regulator.

03 Trusted AI for procurement

At a level below business model structuring and strategy setting, an organization designates which entities are acceptable to interact and conduct business with. These include employees, contractors, service providers, and other entities in a supply chain.

For example, HR outlines principles for employees to meet certain standards and conditions for engagement and disengagement, from where employees can be sourced and where they could be located as employees.

With the phenomenon of digital workers, i.e., "software-based labor,"[6] human resources managers become AI resources managers. They define the roles that these workers can perform in place of human employees, e.g., customer service and complaint resolution. At this level, HR management may rule in or out the ability of these agents to automatically sign contracts.

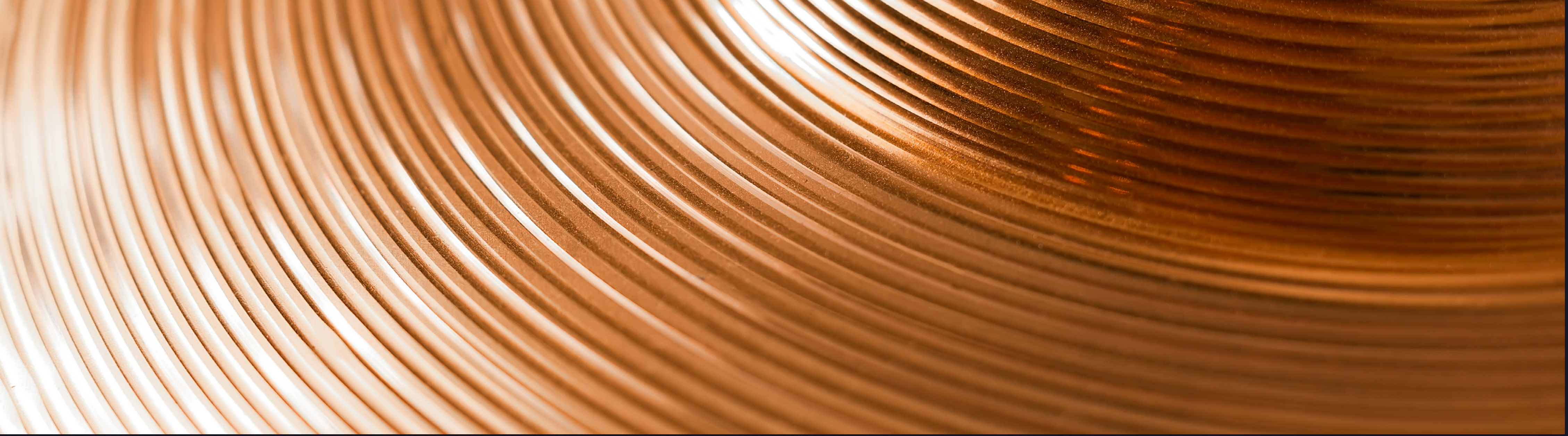
04 Trusted AI for purpose - solution building

At an operational level, ethical AI must be incorporated in design principles and implemented at each operational build phase of data, model and system.

At this level - the solution level of Gen AI operation - one needs to be clear as to purpose when validating inputs to and outputs from models. This means not simply validating what might happen, but being aware of the total scope of what should happen. In practical terms, this means strict exclusion of any ambiguous inputs and outputs.

The context of the activity is important. External contact through a customer relations bot requires controls different from those for an agent connected to an internal knowledge management system, e.g., a customer accessing a customer-focused knowledge base versus an inquiry bot with access to corporate IP responding to technical questions.

[6] IBM, "What is a digital worker?", accessed July 11, 2024



Moving *too fast*

Software development is an industry where generative AI is already assisting with design, development, and testing. However, without specific governance, 63% of software professionals are already using unauthorized tools and exposing their organizations to functional and legal risks.[7]

Functional risks

- *Trust and correctness - no guarantee of the correctness of output*
- *Inherited risk - a foundational/frontier LLM may be built on unknown, unvetted data*
- *Bias - from biased training data*
- *Sustainability - high energy consumption in training and operation*

Legal risks

- *Privacy breaches via user prompts on unvetted training data*
- *Data leakages via user prompts*
- *Copyright breaches revealed in training data not authorized for reproduction*

To gain control of unsanctioned generative AI use, companies must impose a risk assessment framework for software iteration.

Where an organization has a consistent multi-agent AI framework, a software developer will go through training that addresses legal, security and ethical concerns and, in the process, gain certifications for Gen AI tools and prompt engineering.

[7] Capgemini Research Institute, "Turbocharging software with Gen AI", July 2024

Creating a conversational *self-service virtual AI assistant*

A Swedish multinational based in the Netherlands that designs and sells ready-to-assemble furniture requested a generative AI conversational self-service virtual assistant. It had to be capable of interacting with customers for initial inquiries and to resolve issues.

Setting a high bar for trust

A general LLM without specific context training can give misleading responses to customer inquiries. This undermines customers' trust in the chatbot and affects their choice to use it again.

AI models are required by law in many jurisdictions to be responsible, meaning adhering to principles of transparency, fairness, and accountability. Designing a virtual assistant requires attention to any risk that could be caused by an LLM's training data.

Designing and building

To build the chatbot, the project team used GPT 3.5/4 and Llama 7b as the bases for the foundation model. The techniques used included prompt engineering, classification, discriminator models, and fine-tuning.

The model incorporated custom validation to produce trustworthy, human-like, factual responses free from hallucinations, bias, and adversarial speech. It was also instructed on tonality and answer structure.

The virtual assistant made real

The chatbot can understand product attributes, process questions with the right context, and use market-specific content to provide answers. It analyzes customer's questions and phrasing for personalized responses and can handle all pre- and post-sale scenarios. It also augments human workers' responses.

The improved chatbot has significantly increased efficiency and trust in customer service, leading to:

Higher customer satisfaction levels through **tailored responses**

98% accuracy in mitigating hallucinations, bias, and adversarial content

56% more questions *accurately* answered

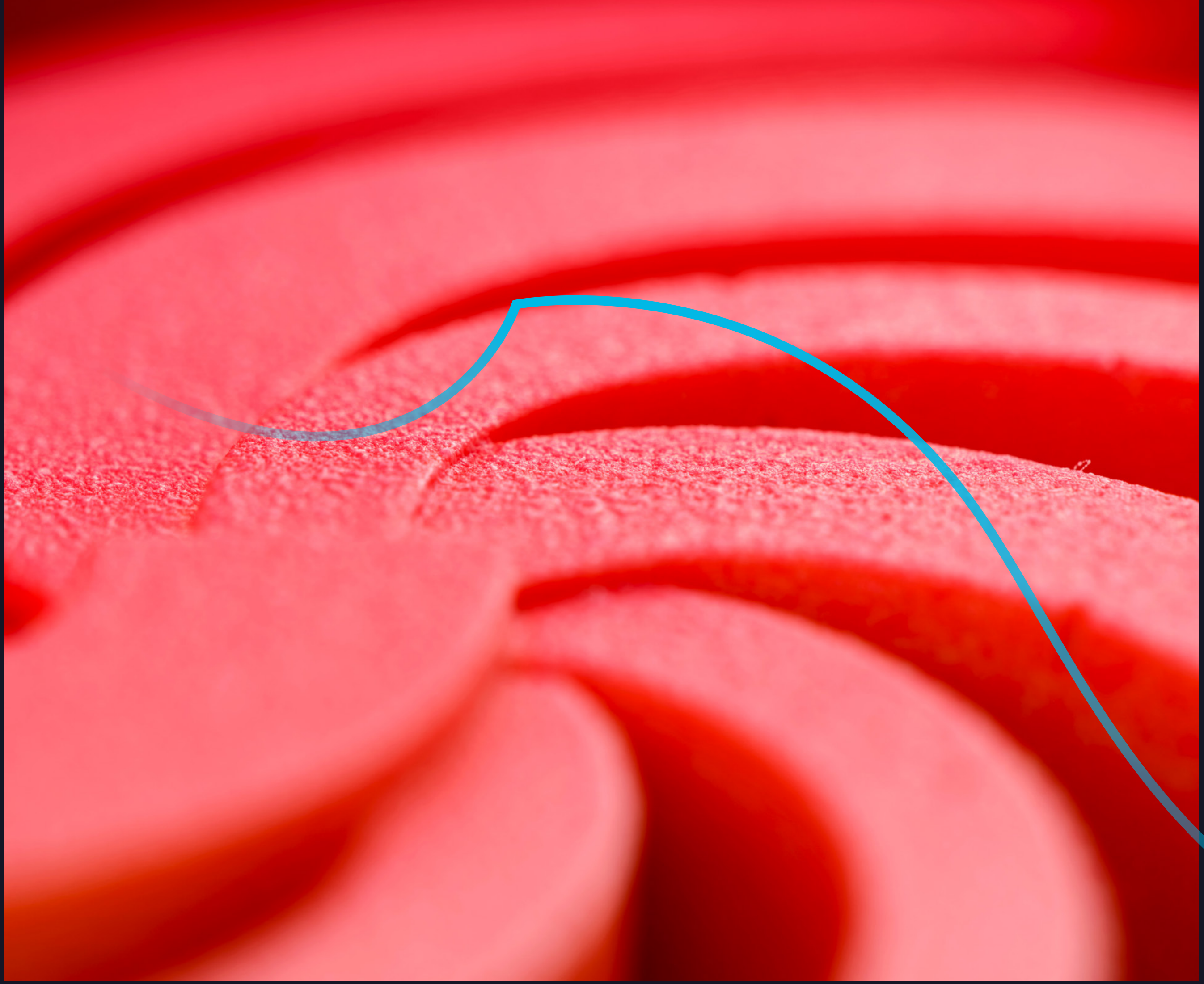
Cost reduction Re-directions to customer support centers *minimized*

Capgemini and *trusted AI*

Capgemini believes that every role within a business has a responsibility to achieve trusted AI. Ensuring that an organization has compliant, ethical and responsible AI can only happen when trusted AI permeates all business governance, technical execution, and operations.

Expert to contact

Steve Jones
Trusted AI Global Lead



About Capgemini

Capgemini is a global business and technology transformation partner, helping organizations to accelerate their dual transition to a digital and sustainable world, while creating tangible impact for enterprises and society. It is a responsible and diverse group of 340,000 team members in more than 50 countries. With its strong over 55-year heritage, Capgemini is trusted by its clients to unlock the value of technology to address the entire breadth of their business needs. It delivers end-to-end services and solutions leveraging strengths from strategy and design to engineering, all fueled by its market leading capabilities in AI, cloud and data, combined with its deep industry expertise and partner ecosystem. The Group reported 2023 global revenues of €22.5 billion.

www.capgemini.com

Capgemini