

A man coding

How trustworthy is your GenAI application?

03 September 2026, 12:00

Mihail Mihaylov

Our Trustworthiness Playbook helps you discover what you need to fully trust your GenAI

Generative AI (GenAI) is rapidly moving from experimental tooling to a core enabling technology in industrial and digital-technology applications. Companies are embedding GenAI into products, workflows, and decision processes to improve efficiency, flexibility, and accessibility. But the real breakthrough for GenAI will depend on the question: “Can you trust your application?” The answer to this question varies by industry, application, and company. Our Trustworthiness Playbook can help you make that assessment.

Different requirements per project and per industry

Trustworthiness does not reside in the AI model alone. The complete workflow, data flow, and process design are equally important. And what trustworthiness requires can differ significantly depending on the context. Treating every possible requirement as equally critical can lead to unnecessary complexity and superfluous engineering efforts.

Consider an internal GenAI assistant that helps a maintenance engineer find information in technical documentation. A wrong answer is undesirable, but the engineer remains in control and can verify it.

Now consider an AI forecasting system whose output is automatically used to submit bids on an energy market. Errors can have direct financial consequences, while noisy or incomplete input data can affect an automated process. Both are GenAI applications. But trusting them requires different priorities.



Nine aspects of trustworthy GenAI

The playbook identifies nine aspects of AI trustworthiness: **accuracy, compliance, explainability, robustness, reliability, safety, transparency, fairness, and latency**.

Accuracy checks whether the system produces the correct output. **Reliability** checks whether it behaves consistently under normal conditions. **Robustness** checks whether it continues to perform when inputs are noisy, incomplete, or different from the data it was designed for.

Other aspects become critical in different circumstances. A system exposed to external users may require greater emphasis on **safety**. In regulated environments, **compliance** and **transparency** may become priorities. **Explainability** matters when users need to understand and validate recommendations. **Fairness** becomes relevant when outputs can affect individuals or groups differently. And in a real-time operator application, even an accurate answer may be useless if **latency** is too high. You do not need to maximise all nine aspects. Focus on the ones your application cannot afford to neglect.

Decision table to guide your considerations

This is where the playbook's **decision table** can play a vital role. Instead of starting from abstract principles, it asks practical questions about your application. A 'yes' answer points towards trustworthiness aspects that should be considered must-haves.

For example:

- Does the AI support important operational, financial, health, or safety decisions? Focus on **safety** and **explainability**
- Is it exposed to external users or untrusted inputs? Prioritise **safety**
- Is input data noisy, drifting, or incomplete? Maximise **robustness**
- Do regulatory or audit requirements apply? Focus on **compliance** and **transparency**
- Must the application respond in real time? Prioritise **low latency**
- Must users or auditors understand its behaviour and limitations? Prioritise **transparency**
- Does it perform prediction, forecasting, automation, or control? **Accuracy and reliability** are your main concerns

This decision table helps you transform a broad discussion about AI risk management into a practical instrument that engineering, product, and innovation teams can actually work with.

Example: interactive operator guidance

The playbook also features concrete examples of how the approach works in practice. Take a production environment, for instance, where employees use an AI-powered assistant to take on more complex tasks independently. The system can combine different sources of information, such as technical documents, images, and spoken questions, and interact with external tools to provide operators with contextual information and recommendations.

As an operator guidance application, **accuracy, reliability, and low latency** are logical starting points. Employees need correct and consistent support, and they expect answers within seconds. But when applying the decision table, additional priorities are revealed.

The assistant generates recommendations and content, making **explainability** an important aspect. The interaction with less tech-savvy operators calls for a focus on **safety** too. And because speech input can be noisy and transcriptions imperfect, **robustness** becomes essential. The system must continue to provide useful support even when its inputs are less than ideal.

The assessment therefore identifies six must-haves: **accuracy, reliability, explainability, safety, robustness, and low latency**. **Transparency and compliance** remain valuable but are not necessarily priorities for this particular application.

These requirements can then translate into practical design choices, such as improving reliability by structuring tool interactions and validation mechanisms, or strengthening safety by restricting the system to predefined actions and keeping operators in the loop.

Conditions for GenAI trustworthiness

A highly intelligent GenAI app may be utterly useless if the right conditions for trustworthiness aren't met. Even the best efforts to make GenAI trustworthy lead to a poor result when input data is poor, outputs are not validated, users misunderstand its limitations, or the workflow gives the AI too much autonomy.

Conversely, there are various ways to improve trustworthiness. Our playbook lists many of these, not as high-level theory, but as practical guidance. Grounding outputs in authoritative data, adding validation checks, standardising prompts, monitoring drift, defining fallback logic, restricting sensitive capabilities, or keeping humans in the loop: these are just a few examples of the practical

advice you can find in this document.

In summary

Trustworthy GenAI requires more than a good AI model. What ‘trustworthy’ means depends on the task, data, users, autonomy, operating environment, consequences of errors, and applicable regulations.

The Sirris and Flanders Make ‘Trustworthiness Playbook’ helps teams make that assessment systematically. It combines nine trustworthiness aspects with recognisable use-case clusters, a practical decision table, and concrete improvement strategies.

Make trustworthiness practical for your GenAI application

Which trustworthiness aspects really matter for your GenAI use case? Download the Trustworthiness Playbook to identify your priorities and discover practical measures to address them.

[Download the playbook](#)

Questions about trustworthy GenAI?

Do you still have questions about trustworthiness or how to apply these principles to your GenAI use case? Mihail Mihaylov, Senior Engineer AI & GenAI at Sirris, will be happy to discuss them with you.

[Contact our expert](#)

Authors

Profile picture for user mihail.mihaylov@sirris..

Mihail Mihaylov