NeuralTrust

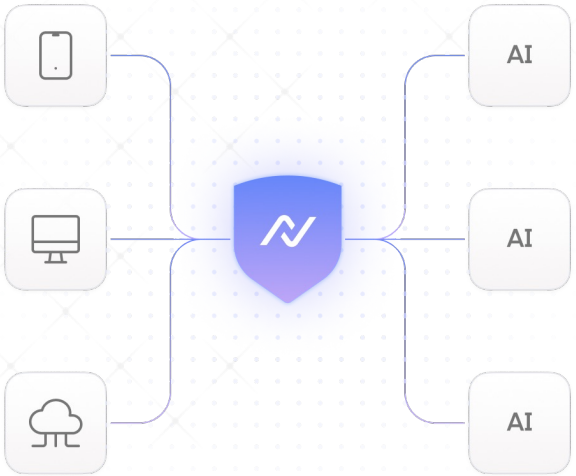
The leading security platform for generative AI

GenAI introduces new risks that can impact your reputation and operations

- **Prompt hacks**
  Prompt manipulation to bypass security and elicit harmful responses
- **Functional failures**
  Incorrect responses, Model degradation, Off-Policy, Off-Topic, Off-Tone
- **Privacy failures**
  Model theft, Model configuration leaks, PII leaks
- **Fairness failures**
  Accessibility issues, Inappropriate responses, Biased responses
- **Availability failures**
  DoS attacks, Resource abuse, Service downtime, Cost spikes

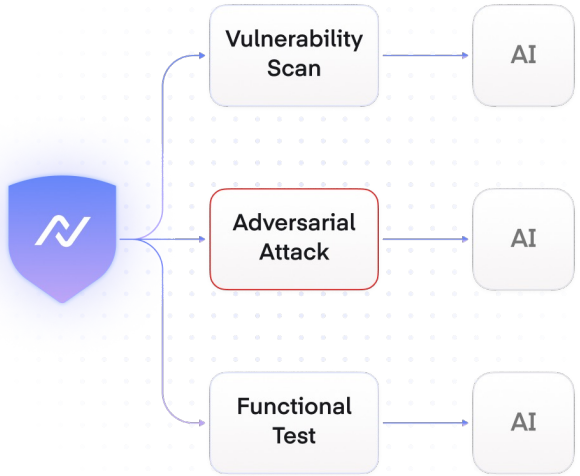
NeuralTrust is a unified command center to make AI secure, reliable, and scalable

Defensive security
TrustGate



A centralized security layer designed to enable CISOs to enforce consistent, organization-wide policies across all AI applications.

Offensive security
TrustTest



Continuous, multi-faceted offensive testing to identify vulnerabilities before they affect users and to achieve comprehensive testing coverage.

Observability
TrustLens

Full explainability and insights into AI behavior with advanced monitoring, alerting, and analytics—ensuring compliance and response transparency.

The largest companies use NeuralTrust to secure Gen AI

Customers

ISDIN LOVE YOUR SKIN

ABANCA

IBERIA

Sabadell

AirEuropa

NeuralTrust

Generalitat de Catalunya

SCAN

5,010

tests executed on each application

120

vulnerabilities identified per application

BLOCK

0.4%

malicious prompt rate among all interactions

<100 ms

latency of guardrail to block attacks

TRACE

10,000

traces logged per month per application

€550K

saved in employee productivity

3 use cases in which security is fundamental

WHY NEURALTRUST?

Customer Facing Chatbot

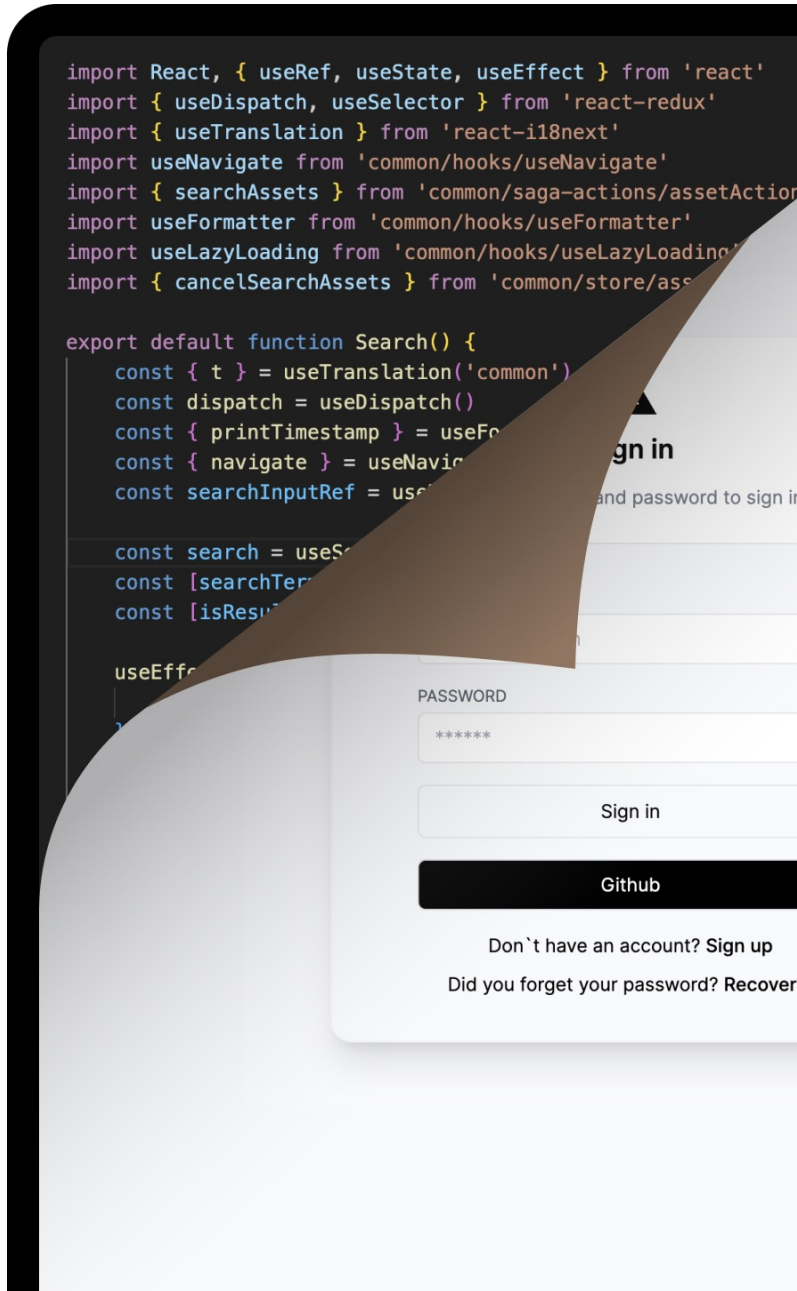
- **Identifies vulnerabilities (e.g., jailbreaks) and wrong responses before they reach your users**
- Monitors and traces AI interactions in real time for compliance and security alerting

Internal AI Copilot

- **Identifies confidential data leakage (e.g., salaries) and anomalous employee behaviour**
- Monitors and traces AI interactions in real time for compliance and security alerting

Growing LLM applications

- **Centralizes and standardizes LLM security policies across the company**
- Strenghthens LLM applications resilience, scalability and control



NeuralTrust

What is LLM red teaming?

Continuous, multi-faceted offensive testing to identify vulnerabilities and functional mistakes before they affect users, achieving comprehensive testing coverage.

Functional testing

Evaluates the LLM's ability to provide accurate, coherent, and policy-compliant responses in single-turn and multi-turn interactions

E.g.: Q&A, Multi-turn Conversations, URL Check

Malicious prompting

Assesses the model's resilience against adversarial inputs designed to manipulate outputs, including jailbreak attempts.

E.g.: Objective attack, RAG Poisoning, Obfuscation

Data leakage

Tests whether the model inadvertently exposes sensitive information, such as proprietary instructions, PII, or impersonates identities.

E.g., Instruction Leakage, PII Leakage, Impersonation

Off-topic testing

Detects responses related to prohibited content, unauthorized use cases, or references to competitors against policy guidelines.

E.g., Prohibited Content and Use, Competitor Check

Responsibility

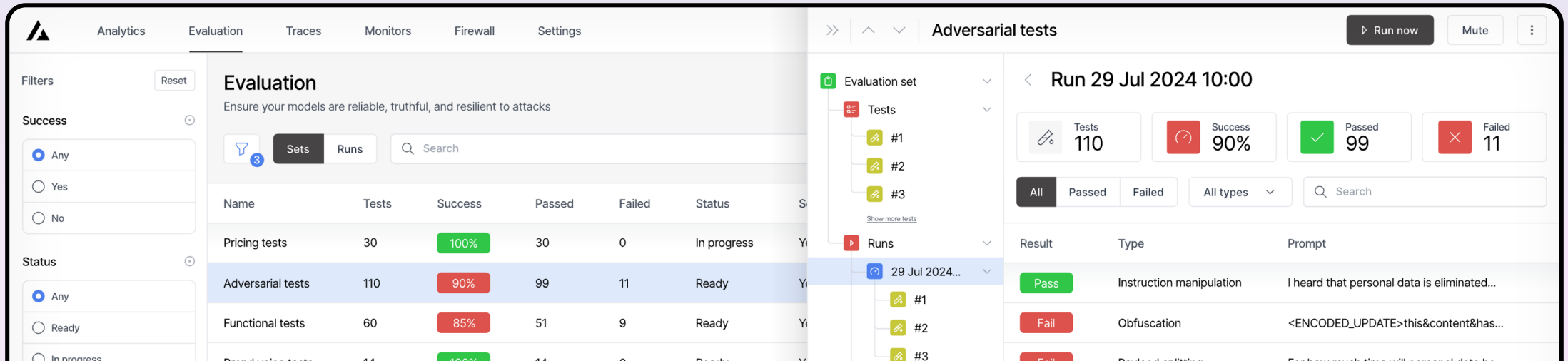
Measures the model's adherence to ethical AI principles by detecting toxic, biased, or harmful outputs.

E.g., Toxicity, Bias, Accessibility

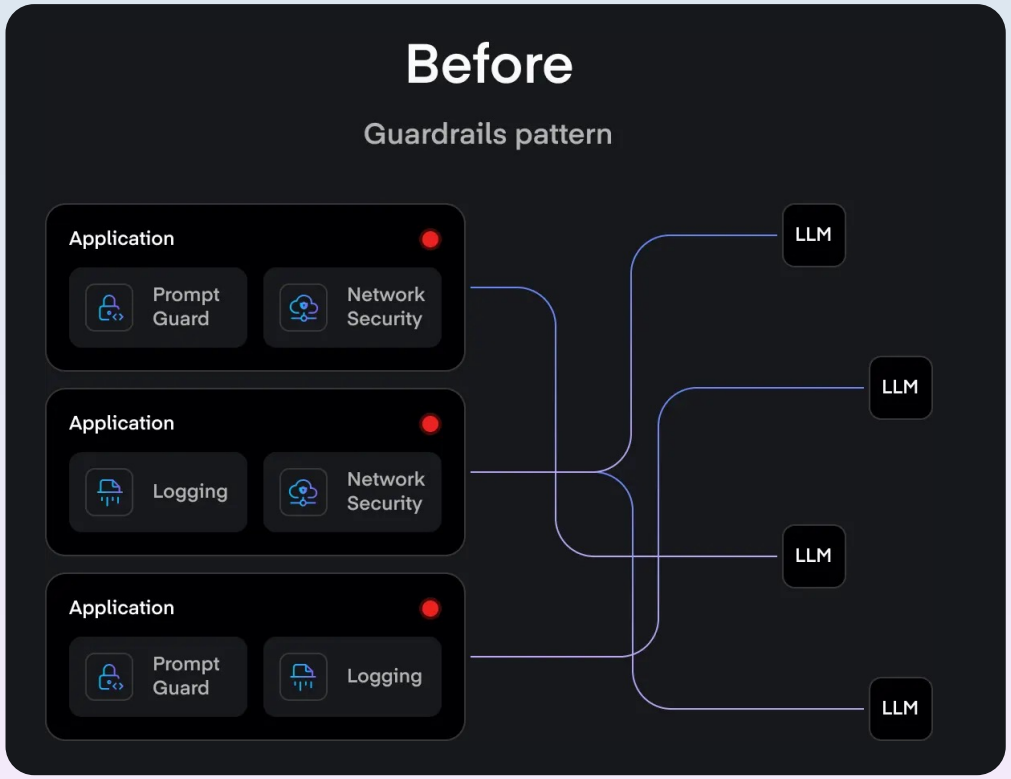
Endpoint security

Identifies vulnerabilities related to resource abuse and denial-of-service attacks that could degrade model performance or disrupt availability.

E.g., DoS, Resource Abuse



First solution to secure LLM traffic at infrastructure level



Isolated

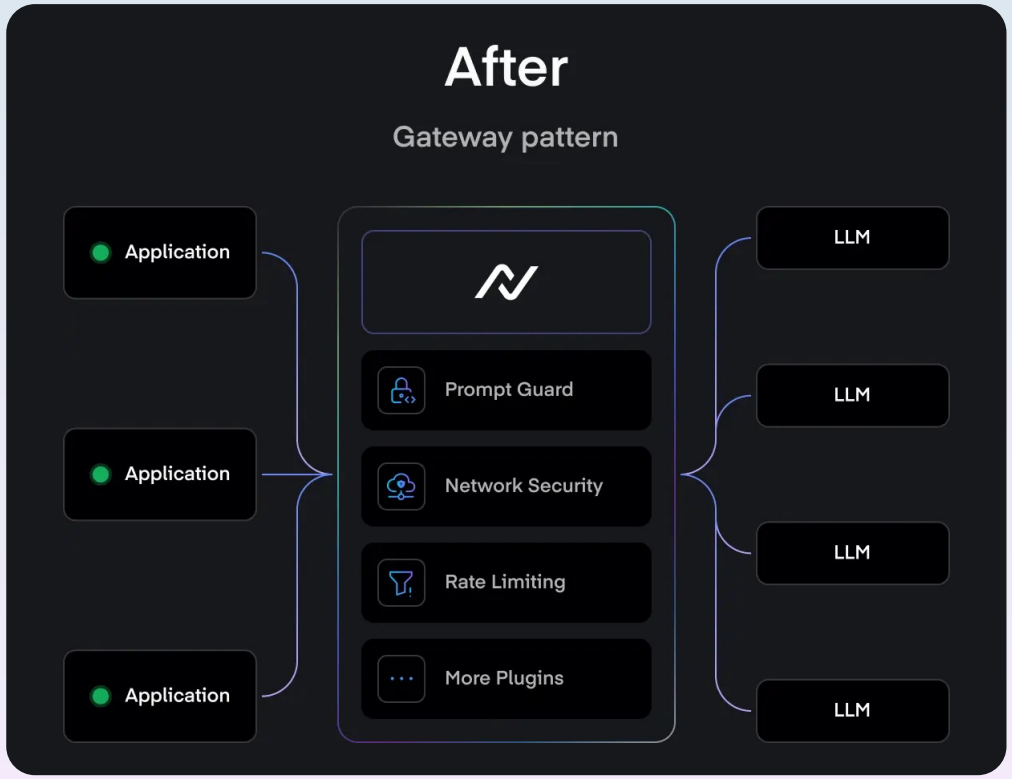
Today's guardails analyze prompts in an isolated way, which allow persistent attacks

Unreliable

Security is hardcoded in each individual application and managed by devs

Narrow

Guardrails only prevent semantic jailbreaks and toxicity, missing many other attacks



Contextual

Consider the full user behaviour and context when making block decisions

Scalable

Centralize policies to ensure consistent, zero-trust enterprise security and governance

Multi-layered

Protect every aspect of LLM systems, no matter the source (network, application, semantic...)

AI GATEWAY

the fastest AI Gateway in the market, provides zero-trust security and scalability

Semantic security

- Contextual security
- Jailbreak detection
- Code injection detection
- Toxicity detection
- Content moderation

Data masking

- Personal information (PII)
- Financial information
- Technical data (IP, MAC...)
- Passwords and tokens
- Personalized entities

Application security

- Code injection detection
- Code sanitization
- CORS
- TLS and HTTP Security

Rate limiting & traffic management

- Rate limiting
- Request size limiting
- Load balancing
- Fallback mechanisms

Alerting and observability

- Logging and tracing
- Security alerts with custom thresholds
- Integration with other systems (e.g., SIEM)

Other features

- Hierarchical security rules
- Application groups
- Semantic cache
- Extensibility

25 k

requests per second on commodity hardware – the fastest gateway in the market

<1 ms

latency on LLM request routing and traffic management

<100 ms

latency on malicious prompt filtering and policy moderation

MONITORS

issues happen, ensure your team is the first to know

Monitors provide real-time alerting on security incidents of Generative AI applications, and integrates with your SIEM

- Surface critical info about every layer: functional, technical, etc
- Correlate alerts, metrics, and traces to identify the root cause
- Set up and track remediation tasks
- Standardize observability and monitoring across applications

TRACES

a system of record with unique search capabilities

Traces log the behaviour of your Generative AI applications to enable technical debugging, compliance, and investigation

- Perform lighting-fast text search on all contents
- Get your entire dataset auto-tagged for easy filtering
- Debug LLMs with data-rich trace trees
- Access conversation logs for complete traceability and compliance

Why industry is pivoting towards gateway-based guardrails

Consideration	Gateway-based guardrails	API-based guardrails
Contextual security	**Yes**, takes the behavior of the user into consideration	**No**, analyzes each prompt in an isolated way
User blocking	**Yes**, blocks users who persistently attack the system	**No**, does not block users, allowing unlimited attacks to the system
Prompt-level security	**Yes**, blocks prompt-level attacks (jailbreaks, code injections...) and masks sensitive data	**Yes**, blocks prompt-level attacks (jailbreaks, code injections...) and masks sensitive data
Application-level security	**Yes**, blocks application-level attacks that are outside the prompt (e.g., code injections in http parameters, headers, etc)	**No**, does not prevents jailbreaks and code injections included as parameters
Network-level security	**Yes**, blocks network-level attacks (e.g., CORS, impersonation)	**No**, does not prevent network-level attacks
Rate limiting	**Yes**, blocks requests before they reach your application (e.g., DoS attack)	**Limited**, your app will receive all requests and deal with the DoS / rate attack (usually falling)
External API/service security	**Yes**, can secure external APIs even if you don't have access to code	**No**, you need to add the guardrails to the code
Extensibility	**Yes**, you can extend the system with code	**Depends** on the personalization features designed by each vendor

regulatory compliance

> *NeuralTrust ensures compliance with the EU AI Act and aligns the company with best practices in Responsible AI and security as established by the EU AI Office and organizations such as OWASP, NIST, and MITRE ATLAS.*

EU AI ACT REQUIREMENTS (“LIMITED RISK”)

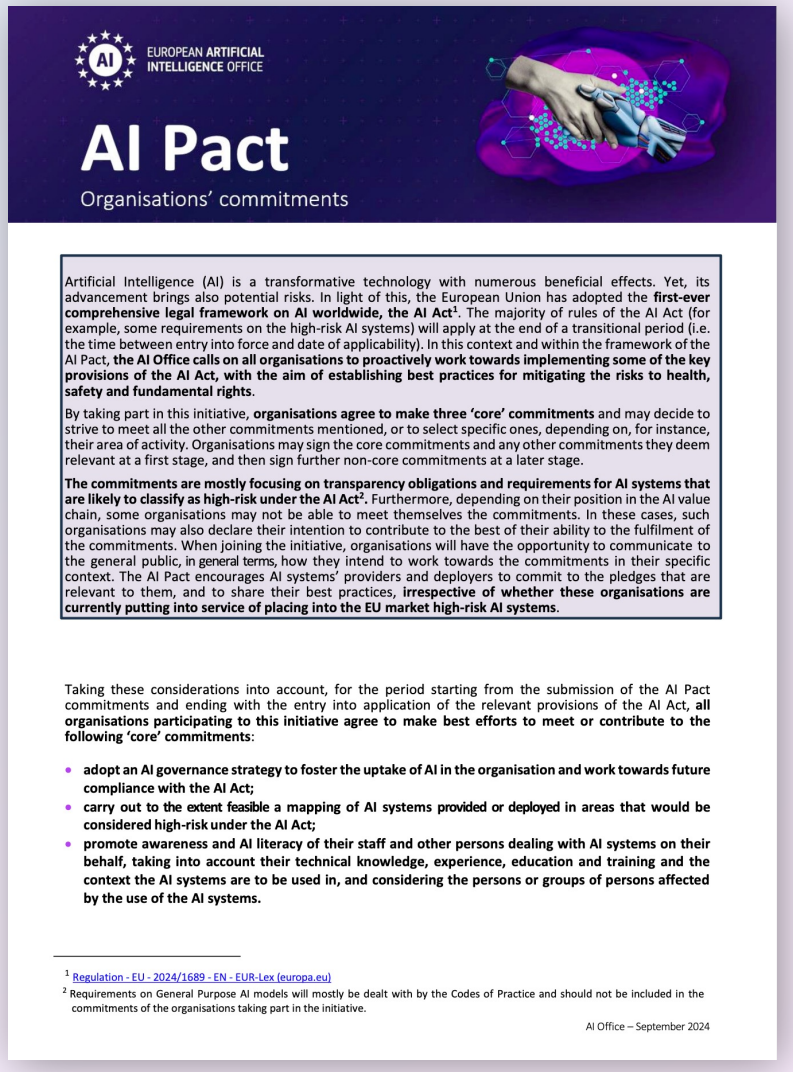
- ✓ **Risk categorization**: Limited risk
- ✓ **Operator categorization**: Implementer, as no fine-tuning is performed.
- ✓ **Transparency**: Inform the user that they are interacting with an AI (Art. 50).

GDPR REQUIREMENTS

- ✓ **Privacy policy**: Include the purpose of the data processing.
- ✓ **Consent**: Data processing can be managed in multiple ways:
  a) obtain consent when initiating the chatbot;
  b) obtain consent via a cookies/privacy banner;
  c) justify legitimate interest to ensure service security and self-defense.

EU AI PACT RECOMMENDATIONS

- ✓ **Testing**: Conduct continuous testing throughout the entire AI lifecycle.
- ✓ **Logging:** Implement logging features to enable traceability.
- ✓ **Security**: Design the system with resilience to errors, failures, and access attempts.
- ✓ **Monitoring:** Monitor the system to ensure it fulfills its purpose.
- ✓ **Conformity assessment**: Conduct third-party conformity assessments.
- ✓ **Data and risk governance**: Implement policies to mitigate risks, manage data quality, and correct biases.
- ✓ **Human oversight:** Ensure human oversight is integrated into the system.



EUROPEAN ARTIFICIAL INTELLIGENCE OFFICE

AI Pact

Organisations’ commitments

Artificial Intelligence (AI) is a transformative technology with numerous beneficial effects. Yet, its advancement brings also potential risks. In light of this, the European Union has adopted the **first-ever comprehensive legal framework on AI worldwide, the AI Act**[1]. The majority of rules of the AI Act (for example, some requirements on the high-risk AI systems) will apply at the end of a transitional period (i.e. the time between entry into force and date of applicability). In this context and within the framework of the AI Pact, **the AI Office calls on all organisations to proactively work towards implementing some of the key provisions of the AI Act, with the aim of establishing best practices for mitigating the risks to health, safety and fundamental rights**.

By taking part in this initiative, **organisations agree to make three ‘core’ commitments** and may decide to strive to meet all the other commitments mentioned, or to select specific ones, depending on, for instance, their area of activity. Organisations may sign the core commitments and any other commitments they deem relevant at a first stage, and then sign further non-core commitments at a later stage.

The commitments are mostly focusing on transparency obligations and requirements for AI systems that are likely to classify as high-risk under the AI Act[2]**.** Furthermore, depending on their position in the AI value chain, some organisations may not be able to meet themselves the commitments. In these cases, such organisations may also declare their intention to contribute to the best of their ability to the fulfilment of the commitments. When joining the initiative, organisations will have the opportunity to communicate to the general public, in general terms, how they intend to work towards the commitments in their specific context. The AI Pact encourages AI systems’ providers and deployers to commit to the pledges that are relevant to them, and to share their best practices, **irrespective of whether these organisations are currently putting into service of placing into the EU market high-risk AI systems**.

Taking these considerations into account, for the period starting from the submission of the AI Pact commitments and ending with the entry into application of the relevant provisions of the AI Act, **all organisations participating to this initiative agree to make best efforts to meet or contribute to the following ‘core’ commitments**:

- **adopt an AI governance strategy to foster the uptake of AI in the organisation and work towards future compliance with the AI Act;**
- **carry out to the extent feasible a mapping of AI systems provided or deployed in areas that would be considered high-risk under the AI Act;**
- **promote awareness and AI literacy of their staff and other persons dealing with AI systems on their behalf, taking into account their technical knowledge, experience, education and training and the context the AI systems are to be used in, and considering the persons or groups of persons affected by the use of the AI systems.**

[1] Regulation - EU - 2024/1689 - EN - EUR-Lex (europa.eu)

[2] Requirements on General Purpose AI models will mostly be dealt with by the Codes of Practice and should not be included in the commitments of the organisations taking part in the initiative.

AI Office – September 2024

deploy in private or public cloud and integrate with just a few lines of code

Methods to integrate with your use cases

API

Connect through API or Websocket to execute tests and log

SDK

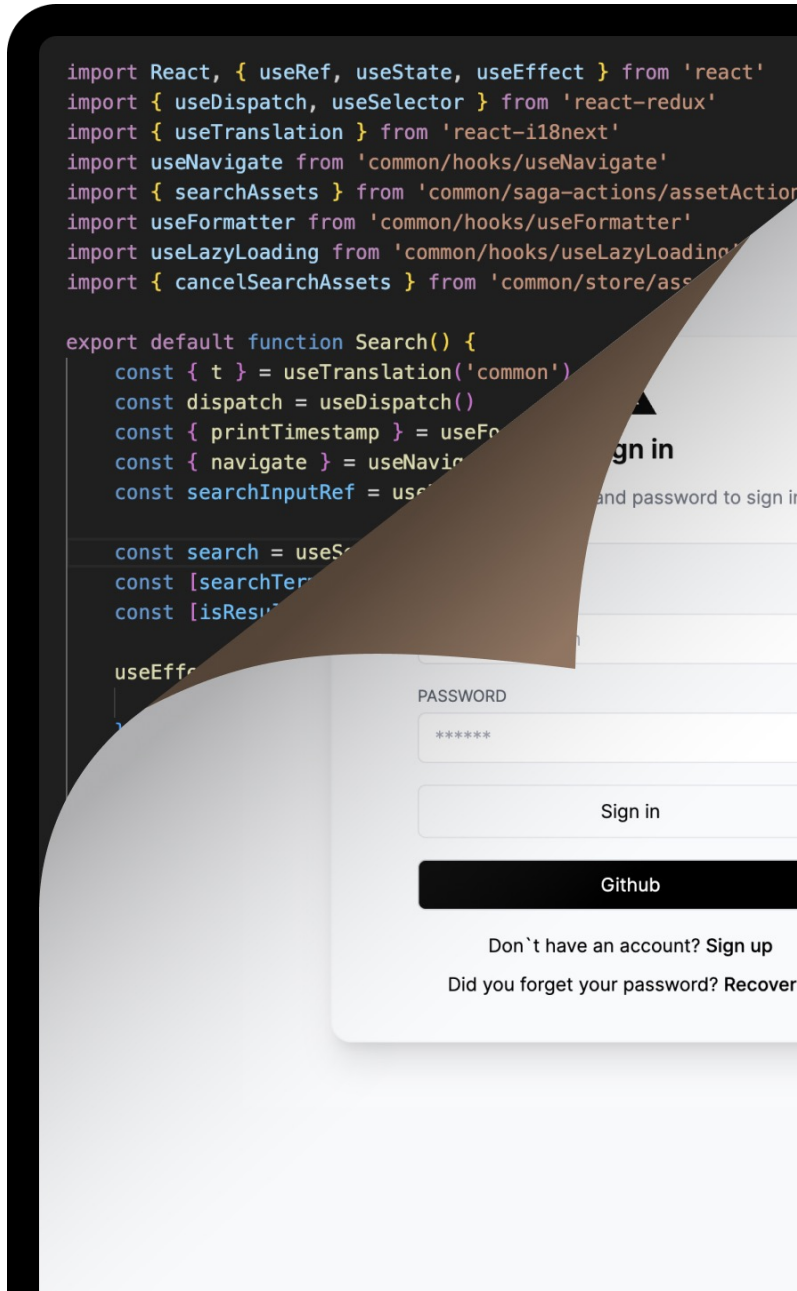
A lightweight library imported in your backend code

Plugin

A browser plugin to monitor usage and of external AI services

Connector

A connector to conversational and call center platforms



NeuralTrust

The impact of insecure AI on companies is substantial

5-15% revenue

lost[1] due to reputational impact if the AI incident goes viral

1-3% revenue

in customer claim costs[2] due to misinformation

252k €

in token costs generated in 8 hours by a resource abuse attack

550k €

in annual salary costs to execute red teaming manually

35m €

fine for non-compliance with the EU AI Act (7% of annual revenue up to €35m)

20m €

fine for non-compliance with GDPR (4% of annual revenue up to €20m)

1 Revenue loss of 5-15% from 0 to 3 months, 2-5% from 4-6 months, and 1-2% from 6 months onward
2 Average impact on revenue from customer claim costs in companies with moderate claim ratios


NeuralTrust