
WHITE PAPERS • 2026 EDITION

Governing AI: from tolerated use to proof of control

Agents, machine identities, data and law: what an executive team must be able to demonstrate in 2026, a vendor-neutral evaluation grid and a 90-day roadmap.

29 October 2026 • 15 min read

KEY POINTS

- According to the 2026 playbook of the French software company Dust, the question for executive teams is no longer whether AI is used, but whether they can prove it is under control.
- Entro Security's report on the first half of 2025 counts 144 non-human identities for every human identity in the environments analysed, up from 92 to 1 a year earlier.
- The “lethal trifecta” described by Simon Willison on 16 June 2025 combines access to private data, exposure to untrusted content and the ability to communicate externally.
- The AI literacy obligation (Article 4 of the AI Act) has applied since 2 February 2025; Regulation (EU) 2026/1744 pushed obligations for Annex III high-risk systems back to 2 December 2027.

- An AI platform should be assessed on evidence: audit trails, admin-only connectors, group-based access (SSO, SCIM), segregation, restricted outbound traffic and reversibility.

In 2024, executive teams wondered whether their staff were using AI. In 2026, the answer is known: yes, often outside the approved tools. So the question has changed. According to the playbook published by the French software company Dust (France · Dust) in a 2026 edition, “shadow AI” is now an accepted fact in large organisations; what an executive committee must now establish is its ability to show that it remains in charge of AI.

This white paper answers that question for executive committees, CIOs, CISOs, DPOs and procurement teams. It draws on that playbook, on the texts in force on 29 October 2026 and on the guidance of European and French authorities. It recommends no product.

1. Why it is urgent: from “shadow AI” to proof of control

Three shifts make proof urgent.

Agents act. A chat assistant produces text. An agent reads a mailbox, queries a database, opens a ticket, sends a message. On 20 July 2026, the French data protection authority and the French Council for AI and Digital Affairs (France · CNIL and CIANum) published an exploratory note on agentic AI. It observes that data flowing between agents, memories and services makes it harder to tell who is responsible for what.

Machine identities are multiplying. Service accounts, API keys, access tokens: every automation has an identity. According to the “NHI & Secrets Risk Report” for the first half of 2025 (Israel · Entro Security), there are 144 non-human identities for every human identity in the enterprise environments the vendor analysed, up from 92 to 1 in the first half of 2024. The study covers more than 27 million identities observed by a

security vendor: it is an order of magnitude, not a representative statistic. Every AI agent adds its own accounts and tokens to that population.

The law asks for evidence. Since 2 February 2025, the AI Act (EU · Regulation (EU) 2024/1689) has required AI literacy measures. The GDPR requires organisations to be able to demonstrate compliance (Article 5(2)). NIS2 and DORA make management accountable for digital risk, including at suppliers. A policy document is no longer enough: you need logs, registers and contracts.

AI governance is no longer judged by the quality of a policy, but by the ability to produce evidence within hours.

2. The five questions an executive team must be able to answer

Dust's playbook groups CISOs' concerns into three families: attribution, data leakage and control. We take them up and add two questions that executive teams underestimate: how tools are rolled out, and the supply chain.

Question 1: attribution

Who did what, and who is accountable? When an agent acts with an employee's credentials, the logs record the employee. Dust's playbook points out that the agent's part then vanishes from the trail: you can no longer tell a human error from an automated action, or which instruction triggered it. The CNIL and CIANum note makes the same point in legal terms: establishing responsibility requires being able to trace every step of the process.

What to demand: a genuine audit trail, linking every action to a person, an agent, that agent's configuration and the data it read. The playbook contrasts this audit trail with mere observability, which counts requests and measures latency without saying who decided what.

Question 2: data leakage and legal exposure

The second concern Dust identifies is data leaving the organisation: towards an external model, an unapproved tool, a third party. The exposure is legal as much as technical: the playbook cites the GDPR in Europe and the CCPA in California (United States · California). For a French organisation, trade secrets, professional secrecy and the specific rules on health data come on top.

Four answers must be written down: which data may go into an AI tool and which never may; whether the provider reuses data to train its models; where data is processed and stored; how long it is kept. Since 2024, the CNIL (France · CNIL) has advised checking to what extent the provider can reuse the data it receives, and adjusting usage accordingly.

Question 3: access control and non-human identities

Dust's playbook sums up control in a few gestures: add, remove, grant, revoke. Then check that the map of access the executive team has in mind matches reality. Its example: making sure that no path leads from the finance team to health data.

Every agent, connector and API key is an identity. Each needs a named owner, a minimal scope, an expiry date and a revocation procedure. With more than a hundred machine identities per human identity, manual review no longer holds: the inventory must be tooled.

Question 4: controlled rollout

The playbook recommends three principles. Connectors, including MCP (Model Context Protocol) servers, tools, secrets and endpoints remain reserved for administrators. Access to agents is granted through directory group membership, with single sign-on (SSO) and automated provisioning (SCIM), not email address by email address. Finally, knowledge boundaries are enforced through restricted spaces.

The benefit is concrete: when an employee leaves or changes role, removal from the directory withdraws their access to every agent, with no manual step.

Question 5: outbound traffic and the supply chain

The playbook recommends restricting outbound traffic to approved addresses only, and notes that remote MCP servers and third-party components are a frequent target of attacks. One case illustrates it. In September 2025, a package imitating the MCP server of an email delivery service added, in its version 1.0.16, a hidden copy of every message to an external address (United Kingdom · The Register).

An MCP server, a downloaded skill, a model, a library or one of the provider’s sub-processors are all links in the chain. Each must be listed, approved and monitored like a supplier.

3. The risk specific to agents: the trifecta and how to break it

On 16 June 2025, the developer Simon Willison (international · Simon Willison) named the central risk of agents: the “lethal trifecta”. An agent becomes dangerous when it combines three capabilities:

INGREDIENT	EXAMPLES	WHAT IT GIVES AN ATTACKER
Access to private data	Email, document management, customer files, HR database	Reach to the information to steal
Exposure to untrusted content	Web page, inbound email, attachment, customer ticket	A way to slip in hidden instructions
Ability to act or communicate externally	Sending email, calling a URL, publishing	A way to get the information out

The mechanism is simple. A language model follows the instructions it reads, including those hidden in a document it is asked to summarise: this is prompt injection. An attacker who plants booby-trapped text in a web page or an email can get internal data sent to an address they control. Willison stresses that vendors' protections do not cover the combinations of tools users assemble themselves, which the MCP protocol actively encourages. He also judges filters insufficient: in security, blocking 95% of attacks is still a failure. His conclusion: avoid the combination.

An agent that combines all three ingredients should be treated as a leak waiting to happen, whatever the provider.

Dust's playbook proposes an architectural answer: separate the roles. A research agent reads the web but has no access to internal data. A drafting agent works on internal data but only receives structured inputs, not free text from outside. A guarded orchestrator alone holds the right to act. Between agents, exchanges go through structured outputs (defined fields, closed lists), which leave little room for a hidden instruction.

The authorities point the same way. The CNIL and CIANum note proposes isolating agents' memory, running them in sandboxed environments, providing an emergency stop and requiring human validation for the most critical decisions. The French national cybersecurity agency (France · ANSSI), in its generative AI recommendations published on 29 April 2024, advises avoiding automated actions between an AI and an application system wherever external data flows in (web, email), and not letting the machine take critical decisions on its own.

A practical rule: for each agent, tick the three boxes. If all three are ticked, the agent does not go into production until it has been split up or placed under systematic human validation.

4. What the law says on 29 October 2026

EU · AI Act

- **AI literacy** (Article 4): since 2 February 2025, providers and deployers have taken measures to ensure a sufficient level of AI literacy among their staff and the people using AI on their behalf. Regulation (EU) 2026/1744, the “AI Omnibus”, which entered into force on 27 July 2026, reworded the article: it now requires measures to support the development of that literacy, without guaranteeing any individual level. The obligation remains; the Commission and Member States must support organisations’ efforts.
- **Prohibited practices** (Article 5): since 2 February 2025. For an employer, the most tangible one covers emotion recognition in the workplace, banned except for medical or safety reasons.
- **Transparency** (Article 50): since 2 August 2026, deployers must disclose deepfakes, as well as generated text published to inform the public on matters of public interest, unless it has undergone human review under editorial responsibility. Providers of generative tools already on the market have until 2 December 2026 to mark the content they produce.
- **High risk**: following Regulation (EU) 2026/1744, obligations apply from 2 December 2027 for Annex III systems (recruitment, worker management, credit, education, access to essential services...) and from 2 August 2028 for Annex I systems built into regulated products.
- **Obligations of deployers of high-risk systems** (Article 26): use the system in line with its instructions, assign human oversight to competent people with the necessary authority, monitor its operation, keep logs for at least six months, and inform workers’ representatives and affected workers before putting it into service at work. Public bodies and some private deployers, notably in credit and in life or health insurance, must also carry out a fundamental rights impact assessment (Article 27).

EU · GDPR

- **Lawful basis** (Article 6) and a defined purpose for every use involving personal data.

- **Data protection impact assessment** (Article 35) where processing is likely to result in a high risk: this is common as soon as AI concerns employees, customers or sensitive data. The AI Act provides that deployers use the provider's information to carry it out (Article 26(9)).
- **Automated decisions** (Article 22): individuals have the right not to be subject to a decision based solely on automated processing that significantly affects them. The CNIL and CIANum note recalls that a purely formal human sign-off is not enough.
- **Processors** (Article 28): written contract, sub-processors subject to prior authorisation, audit rights, assistance in the event of a breach.
- **Transfers** (Chapter V): outside the European Economic Area, an adequacy decision or appropriate safeguards such as standard contractual clauses.

EU · NIS2 and DORA

- **NIS2** (Directive (EU) 2022/2555): for essential and important entities, management bodies approve risk-management measures and can be held liable (Article 20). The measures include supply chain security, access control and multi-factor authentication (Article 21). In France, the transposing law had not been finally adopted by 29 October 2026.
- **DORA** (Regulation (EU) 2022/2554): applicable to financial entities since 17 January 2025. An AI platform is an ICT third-party service provider: it goes into the register of information on contractual arrangements, and its contract must contain the clauses of Article 30 (data location, access to and return of data, termination), plus audit and exit strategy if it supports a critical or important function.

France · CNIL

- **Generative AI**: since July 2024, the CNIL's questions and answers have asked organisations deploying a generative AI tool to start from a precise need, frame usage with an internal policy, prefer on-premise deployment for sensitive data, train users and involve the DPO, working with the CISO.

- **Agentic AI**: the note of 20 July 2026, written with CIANum, holds that agents remain fully subject to the GDPR and proposes technical safeguards (isolation, sandboxed environments, emergency stop, human validation).

Voluntary standards

- **ISO/IEC 42001:2023** (international · ISO and IEC): the certifiable AI management system standard, built on the same structure as ISO/IEC 27001. An ISO/IEC 42001 certificate does not, on its own, give a presumption of conformity with the AI Act; that role belongs to European harmonised standards, the first of which was cited in the Official Journal in October 2026.
- **AI RMF** (United States · NIST): a framework published on 26 January 2023, organised around four functions (govern, map, measure, manage), supplemented on 26 July 2024 by a generative AI profile. NIST states that it is being revised. It has no legal force in Europe, but gives a common vocabulary with US suppliers.

France · hosting sensitive data

An order of 12 August 2026, published in the Official Journal of 14 August, approves version 3.2 of ANSSI's SecNumCloud framework for data of particular sensitivity held by the State, its operators and certain public interest groupings. Compliance is shown by an ANSSI qualification or by a European certification recognised as equivalent. Private companies are not bound by it, but the framework is a benchmark for data exposed to non-European laws. In its 2024 guide, ANSSI recommends preferring qualified hosting when a generative AI system is deployed in the cloud.

Pushing high-risk obligations back to December 2027 does not push back AI literacy, transparency or the GDPR.

5. Data, the foundation of governance

An AI platform is only as good as the data it reaches and the rights that protect it. Six workstreams condition everything else.

Classification. Four levels are often enough: public, internal, confidential, regulated (health data, banking data, protected secrets). Each level comes with a usage rule for each category of tool.

Register. The GDPR record of processing activities (Article 30) can hold AI systems: purpose, data, agents, connectors, suppliers, retention period. A single register, reviewed at each change, avoids two inventories drifting apart.

Inherited access rights. An assistant that queries internal documents, using the technique known as RAG (retrieval-augmented generation), must apply the source documents' permissions to every question. Otherwise, any user can obtain in one sentence the summary of a note reserved for the executive team. This inheritance is tested with several user profiles before opening.

Retention. Prompts, answers, logs and agents' memory are data. Each needs a retention period. The CNIL and CIANum note warns that agents' persistent memory tends to accumulate data and build very precise user profiles.

Hosting and sovereignty. Place of processing, place of storage, sub-processors, and the law applicable to the provider and its parent company: these four pieces of information belong in the procurement file. For health data in France, HDS certification is required on top.

Reversibility. Since 12 September 2025, the Data Act (EU · Regulation (EU) 2023/2854) has governed switching between data processing service providers; from 12 January 2027, switching and data egress charges will be banned. For an AI platform, reversibility also covers agents' instructions, knowledge bases, histories and logs.

6. Evaluation grid for an AI platform

This grid is vendor-neutral. It applies to an off-the-shelf assistant, an agent platform or an in-house build.

CRITERION	EVIDENCE TO REQUEST	WHERE TO FIND IT
Audit trails	A log extract for one agent action: requester, agent, version, rights used, data read, action, timestamp; retention period; export to your monitoring tools	Demonstration on a test environment; admin documentation; security schedule of the contract
Connectors under control	List of connectors, MCP servers and tools that can be enabled; who can add them; approval workflow; token scope	Admin console; vendor security policy
SSO and SCIM	Single sign-on on your directory; group-based access; automatic deprovisioning	Integration documentation; test with a fictitious leaver
Segregation	Restricted spaces; inheritance of source permissions; separation between agents	Demonstration with several profiles; penetration test report
Outbound traffic	List of outbound destinations; configurable allow-list; blocking of external links and images in answers	Architecture diagram; admin settings
Model choice and no lock-in	Available models; ability to switch, or to host one in Europe or on-premise; commitment not to train on your data	Contract; model providers' terms
Hosting and sub-processors	Places of processing and storage; list of sub-processors; exposure to non-European laws; qualifications (SecNumCloud, HDS)	Published sub-processor list; certificates; vendor trust centre
GDPR processing agreement	Article 28 clauses; breach notification; audit rights; transfer safeguards	Data processing agreement annexed to the contract
Reversibility	Export of agents, instructions, knowledge bases, histories and logs in an open format; timeframe; cost	Contractual reversibility plan; partial exit test

CRITERION	EVIDENCE TO REQUEST	WHERE TO FIND IT
Costs	Pricing model (per seat, usage-based); caps and alerts; three-year total cost including integration	Price list; simulation on your actual volumes

Evidence that only exists in a sales deck is not evidence. Ask for the demonstration, the extract, the clause.

7. A 90-day roadmap

Days 1 to 30: inventory, rules, roles

- **Inventory:** AI tools in use, approved or not; agents; connectors; non-human identities; AI features added to your existing software.
- **Acceptable use policy:** two pages at most. Data allowed per classification level, approved tools, prohibited uses, cases where human validation is mandatory.
- **Roles:** an executive sponsor who arbitrates; the CISO, for identities, logs and tests; the DPO, for lawful basis, impact assessment and register; a business owner for each use case; procurement, for the grid, clauses and reversibility.
- **AI literacy:** a first training module, tailored to roles, with attendance recorded.

Days 31 to 60: a governed pilot

- Two or three use cases, on a single platform assessed with the grid.
- Group-based access (SSO, SCIM), approved connectors, restricted spaces.
- A trifecta check for each agent; splitting where needed.
- Impact assessment where the use case requires it; logs enabled and exported.
- Weekly indicators: usage, answer quality, incidents, costs.

Days 61 to 90: control, evidence, scaling up

- **Evidence drill:** pick three agent actions at random and reconstruct, within twenty-four hours, who asked for what, with which rights, on which data.
- **Tests:** a fictitious leaver, booby-trapped content submitted to an agent, a partial data export.
- **Evidence file:** register, impact assessments, contracts, log extracts, training records.
- **Decision:** extend, correct or stop, presented to the executive committee with its indicators.

Scaling up is decided on an evidence file, not on an adoption rate.

8. Ten questions to ask on Monday morning

1. Can we list every AI tool and agent used in the organisation, including outside approved tools?
2. For an action an agent carried out yesterday, can we say who requested it, with which rights and on which data?
3. How many non-human identities do we have, and does each one have an owner and an expiry date?
4. Who can add a connector, an MCP server or a skill, and who approves it?
5. Does a departing employee automatically lose access to every agent and its data?
6. Which of our agents combine all three ingredients of the trifecta?
7. To which destinations can our agents send data?
8. Which suppliers and sub-processors handle our prompts and documents, and in which country?
9. Could we switch model or platform in under three months, with our agents and histories?

10. What AI literacy measures have we taken under Article 4 of the AI Act, and can we prove it?

An executive team that answers these ten questions with documents rather than intentions is governing its AI.

Frequently asked questions

What is the “lethal trifecta” of AI agents? +

Has the AI literacy obligation in Article 4 of the AI Act been removed? +

When do the obligations for high-risk AI systems apply? +

Is ISO/IEC 42001 certification enough to comply with the AI Act? +

SOURCES

1. France · Dust — La gouvernance de l'IA : concilier performance et sécurité dans les grandes entreprises (playbook, édition 2026) (<https://dust.tt>)
2. International · Simon Willison — The lethal trifecta for AI agents: private data, untrusted content, and external communication (16 juin 2025) (<https://simonwillison.net/2025/Jun/16/the-lethal-trifecta/>)
3. Israël · Entro Security — NHI & Secrets Risk Report, premier semestre 2025 (<https://lp.entro.security/nhi-secrets-risk-report-h1-2025>)
4. UE · Journal officiel — Règlement (UE) 2024/1689 sur l'intelligence artificielle, AI Act (13 juin 2024) (<https://eur-lex.europa.eu/eli/reg/2024/1689/oj>)
5. UE · Journal officiel — Règlement (UE) 2026/1744, omnibus numérique sur l'IA (24 juillet 2026) (https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202601744)
6. UE · Journal officiel — Règlement (UE) 2016/679, RGPD (27 avril 2016) (<https://eur-lex.europa.eu/eli/reg/2016/679/oj>)
7. UE · Journal officiel — Directive (UE) 2022/2555, NIS2 (14 décembre 2022) (<https://eur-lex.europa.eu/eli/dir/2022/2555/oj>)

8. UE · Journal officiel — Règlement (UE) 2022/2554, DORA (14 décembre 2022) (<https://eur-lex.europa.eu/eli/reg/2022/2554/oj>)
9. France · CNIL — IA agentique et données personnelles : note exploratoire avec le Conseil de l'IA et du numérique (20 juillet 2026) (<https://www.cnil.fr/fr/ia-agentique-cnil-cianum-note>)
10. France · CNIL — Les questions-réponses de la CNIL sur l'utilisation d'un système d'IA générative (18 juillet 2024) (<https://www.cnil.fr/fr/les-questions-reponses-de-la-cnil-sur-lutilisation-dun-systeme-dia-generative>)
11. France · ANSSI — Recommandations de sécurité pour un système d'IA générative (29 avril 2024) (https://cyber.gouv.fr/sites/default/files/document/Recommandations_de_s%C3%A9curit%C3%A9_pour_un_syst%C3%A8me_d_IA_g%C3%A9n%C3%A9rative.pdf)
12. France · Légifrance — Arrêté du 12 août 2026 portant approbation du référentiel d'exigences SecNumCloud (JO du 14 août 2026) (<https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000054678082>)
13. International · ISO et IEC — ISO/IEC 42001:2023, système de management de l'intelligence artificielle (<https://www.iso.org/standard/42001>)
14. États-Unis · NIST — AI Risk Management Framework, AI RMF 1.0 (26 janvier 2023) et profil IA générative (26 juillet 2024) (<https://www.nist.gov/itl/ai-risk-management-framework>)
15. Royaume-Uni · The Register — Postmark MCP server code hijacked : copie cachée des e-mails (29 septembre 2025) (https://www.theregister.com/2025/09/29/postmark_mcp_server_code_hijacked/)

This white paper is independent analysis, not legal advice. Facts checked at the edition date; every source is cited.