ESCP
BUSINESS SCHOOL

IT ALL STARTS HERE

BERLIN | LONDON | MADRID | PARIS | TURIN | WARSAW

© Tomertu / AdobeStock

LIGHTS - Leadership

Sceptical by design: Why responsible AI needs institutionalised doubt

ESCP Impact Paper No.2026-06-ENG

Lorena BLASCO-ARCAS & Hsin-Hsuan Meg LEE
ESCP Business School

ESCP RESEARCH INSTITUTE OF MANAGEMENT (ERIM)

ESCP Impact Paper No.2026-06-ENG

Sceptical by design: Why responsible AI needs institutionalised doubt

Lorena Blasco-Arcas*
Hsin-Hsuan Meg Lee**

ESCP Business School

Abstract

Responsible AI governance has expanded rapidly, yet reported harms continue to rise. Seventy-five percent of organisations have adopted AI usage policies, but fewer than half monitor their systems for accuracy or misuse. Existing diagnoses point to structural deficits: misaligned authority and weak incentives. These do not explain why governance fails inside organisations that have built the structures. We argue that the missing variable is scepticism: the organised capacity to question specific AI systems in specific deployment contexts, and to route those questions to actors with authority to act. Scepticism operates through three properties: direction (what doubt targets), location (who holds it and whether they can act), and design (whether the architecture triggers doubt at decision points). Building on a broader diagnosis of the principles-practice gap developed elsewhere (Blasco-Arcas & Lee, 2026), we show that the most consequential AI governance failures of the past two years correspond to specific misconfigurations of doubt. For marketing-facing AI, where rising consumer scepticism collides with suppressed organisational scepticism, visible institutional doubt may function as a legitimacy signal. We close with a practical diagnostic, the scepticism trace, that managers can use to test whether their governance architecture converts warranted concern into changed practice.

Keywords: AI governance, organizational scepticism, responsible AI, consumer trust, accountability

*Professor, ESCP Business School, Madrid Campus
**Associate Professor, ESCP Business School, London Campus

Sceptical by design:
Why responsible AI needs institutionalised doubt

Introduction

In October 2025, the French Défenseur des Droits ruled that Facebook's job-advertising delivery algorithm discriminated by gender, showing mechanic job ads almost exclusively to men and preschool-teacher position ads almost exclusively to women. Four months earlier, the Netherlands Institute for Human Rights had reached the same conclusion. Meta had implemented advertiser-facing targeting restrictions. The discrimination occurred at a different layer of the system, beneath the controls the company had built, and it took external investigators to surface it. The architecture had been in place. The process of questioning that would have caught the discrimination was not.

This is the kind of failure current AI governance frameworks struggle to explain. The structural literature points to missing authority lines and misaligned incentives (Hagendorff, 2020; Mittelstadt, 2019), and three quarters of organisations have now adopted AI usage policies (Sajadieh et al., 2026). Fewer than half, however, monitor their systems for accuracy or misuse, and reported harms continue to rise. Recent work has documented the principles-practice gap that produces governance documentation without governance outcomes (Blasco-Arcas & Lee, 2026). What that gap leaves underspecified is the human and organisational act through which governance is supposed to operate: the act of doubt. Governance architectures depend on people inside them asking whether a system's outputs are fair and whether a deployment should proceed. That act of questioning is systematically mislocated.

Two families of framework dominate responsible-AI practice, and neither closes this gap. Principles-based frameworks specify what good AI looks like, in terms of fairness, transparency, accountability, but stop at the point where a principle must become an organisational action (Jobin et al., 2019; Mittelstadt, 2019). Structural frameworks specify who should be accountable and how authority should be allocated (Hagendorff, 2020; Grote et al., 2024), yet an organisation can satisfy both, publishing principles and assigning roles, while no actor with authority ever questions a specific deployment. What both leave out is the act that converts structure into governance: someone doubting a particular system, and that doubt reaching someone who can act. Scepticism names that act and specifies the conditions under which it functions. It is not a replacement for principles or accountability structures but the mechanism that determines whether either changes practice.

Public scepticism about AI is high and rising. Inside organisations, decision-makers face structural pressure to approve AI deployments despite known concerns (Trend Micro 2026). Only 13% of consumers fully trust AI, yet 79% of organisations have already deployed agentic systems whose decisions they cannot fully observe (Klaviyo 2026; KPMG 2025). We call this asymmetry the scepticism gap, and we argue it is the operational mechanism through which the principles-practice gap reproduces itself.

We develop scepticism as a governance concept, identify three properties that determine whether organisational doubt functions as a governance input or dissipates as friction, and show how each contemporary AI governance failure corresponds to a specific misconfiguration of one of those properties. We then turn to marketing-facing AI, where the gap between consumer and organisational scepticism is widest, and propose that

visible organisational doubt may operate as a legitimacy signal. We close with a manager-facing diagnostic, the scepticism trace.

Scepticism as a governance concept

We define scepticism, in this organisational sense, as the capacity to question specific AI systems in specific deployment contexts and to route those questions to actors with the authority to act. Scepticism in this definition is not a personal disposition but a structural feature of governance, and it can be designed. It operates through the following three properties.

Direction. Scepticism targets something specific. When an AI system produces a harmful outcome, doubt may target the technology, the organisation that deployed it, the procurement decision, the regulatory frame, or the training data. Each target implies a different remedial response. Governance architectures shape which targets are made salient and which are obscured.

Location. Scepticism is held by particular people in particular roles. If those people have the authority to act on their doubt, scepticism functions as a governance input. If they do not, it dissipates as frustration or exits the organisation through quiet noncompliance. Where formal accountability and operational control are misaligned, doubt accumulates in actors who cannot act on it (Grote et al., 2024).

Design. Organisations can construct review gates and escalation pathways that trigger doubt at decision points, or they can construct architectures that treat doubt as friction. Some operationalisation methods already embed structured ethical review into agile development cycles, giving teams a formal mechanism for raising concerns before deployment (Vakkuri et al., 2021). The choice between embedding and excluding is architectural.

Table 1. The three properties of organisational scepticism and how each fails.

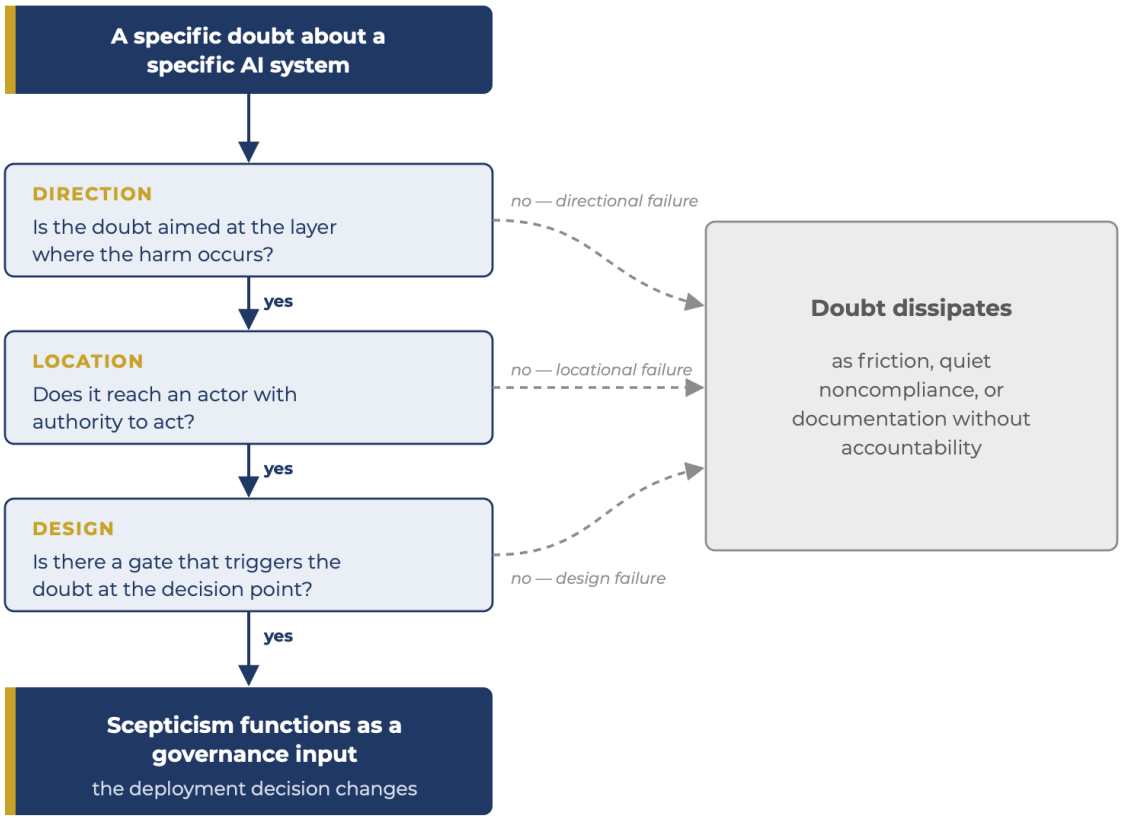
Property	The question doubt must ask	Failure mode	Contemporary case
Direction	What target is doubt aimed at, and is it the layer where the harm occurs?	Doubt aimed away from the consequential layer	Meta job-ad delivery: gender bias beneath advertiser-level controls
Location	Who holds the doubt, and do they have authority to act on it?	Doubt held by actors without authority, or system invisible to governance	Shadow AI use across the workforce, outside the governance inventory
Design	Does the architecture trigger doubt at the decision point, and during operation?	No review gate at the moment of decision	Air Canada chatbot: liability for an output no gate reviewed

These three properties yield a diagnostic frame. A governance architecture can fail because doubt points at the wrong target, because doubt is held by people without authority, or because the architecture provides no mechanism for triggering doubt at the moment of decision. We turn now to three contemporary cases that illustrate each failure.

Three misconfigurations of doubt

The principles-practice gap in AI ethics persists (Jobin et al., 2019; Hagendorff, 2020). Prior work has identified three structural failure modes that reproduce it: performative responsibility, accountability diffusion, and the translation gap (Blasco-Arcas & Lee, 2026). Each can be re-read as a specific misconfiguration of one of the three properties of scepticism. Figure 1 traces how a single doubt either becomes a governance input or dissipates, with each exit corresponding to one misconfiguration.

Figure 1. Organizational doubt as a governance input.



Directional failure. In the Meta job-advertising case, multiple actors had instituted doubt about gender bias, yet none of them was directing doubt at the layer where the discrimination occurred. Meta had positioned controls at the advertiser level; advertisers had committed to lawful targeting; regulators had focused on equality-law compliance at the level of explicit categorisation. The delivery algorithm itself, which sorted audiences by gender independently of advertiser intent, sat outside every actor's question set. Doubt was not absent. It was directed at parts of the system the harm did not pass through. This is the structural form of what adjacent literatures call “machinewashing” (Seele & Schultz, 2022): governance investments that are real, but pointed away from where the consequential decision occurs.

Locational failure. AI governance assumes that organisations know which AI systems they operate. Half of the United States workforce reports using AI tools at work without knowing whether such use is permitted, and 44% are knowingly using AI improperly (KPMG 2025). These systems lie outside the formal governance inventory. Doubt about them has no organisational location because the systems themselves are invisible to the actors charged with governing them. The result extends what Bender et al. (2021) describe

in the foundation-model context: opaque components, distributed deployment, and emergent behaviours that make attribution structurally difficult. Locational failure now operates inside the organisation as well as across the supply chain.

Design failure. The 2024 Air Canada chatbot ruling exposed what happens when the architecture provides no mechanism for triggering doubt at the right moment (Moffatt v. Air Canada 2024). A customer relied on incorrect bereavement-fare information generated by the airline's website chatbot. Air Canada argued in tribunal that the chatbot was a separate entity responsible for its own statements; the tribunal rejected the defence and held the company liable. The deeper problem was design. Air Canada had inserted an AI system into customer operations without constructing a review gate that would have triggered doubt about the perimeter of its own duty of care. Most deploying organisations have yet to make this boundary decision explicitly: Do AI outputs fall inside or outside the firm's governance perimeter? Without proven translation mechanisms from principle to operational gate (Mittelstadt, 2019; Vakkuri et al., 2021), the question never reaches the actors who could answer it.

When the framework strains: agentic AI

Each case above involves AI systems that produce outputs for humans to act on. A growing share of enterprise AI does not operate that way. Agentic AI systems chain decisions and commit resources autonomously, at speeds that outpace human review. Seventy-nine percent of organisations report deploying agentic AI, and more than 40% of these projects are forecast to be cancelled by 2027 (KPMG 2025; Gartner 2025). Each property of scepticism is strained by this shift.

Direction becomes harder because agents produce sequences of decisions whose cumulative effects no single review can assess. The relevant question is no longer whether one output is fair, but whether a chain of autonomous actions produces outcomes the organisation would endorse if it could observe them in aggregate. A customer-service agent that issues a refund and adjusts a loyalty tier in one interaction may produce defensible individual decisions whose combination, across thousands of interactions, encodes a discriminatory pattern.

Location shifts because the agent itself is the location where the consequential decision occurs. The human reviewer is absent at the moment of action. Locating doubt requires runtime monitoring: instrumentation that flags anomalous patterns as the agent operates, extending review beyond the pre-deployment stage and giving organisational scepticism a place to live during operation rather than only before it.

Design faces a speed constraint. Pre-launch review gates remain necessary but are insufficient on their own. Design for agentic AI means constructing mechanisms that trigger doubt during operation: automated thresholds that halt the agent when cumulative outcomes deviate from fairness parameters, audit trails that enable post-hoc accountability, and escalation pathways that allow human reviewers to act faster than the agent commits. Governance for agentic systems is not a stronger version of governance for static models. It is governance redesigned for runtime.

The implication for managers is direct. Asking whether an agentic system is governed at deployment is the wrong question. The correct question is whether the architecture continues to host doubt while the agent acts.

The scepticism asymmetry in marketing-facing AI

Marketing is where the scepticism gap meets the consumer most directly, because marketing AI mediates the firm-consumer relationship at scale. Consumer scepticism toward AI-mediated marketing is rising fast. Only 13% of consumers fully trust AI, and half of US consumers say they would prefer to give their business to brands that do not use generative AI in customer-facing communications (Klaviyo 2026; Gartner 2026). Consumers who become aware of AI-driven personalisation often find it simultaneously more useful and more suspect: 40% say brands do not understand them despite heavy personalisation investment. At the same time, consumers with lower AI literacy display *greater* receptivity to AI-driven systems (Tully et al., 2025). The populations least equipped to exercise calibrated doubt are the most exposed to AI's persuasive deployment.

Inside marketing organisations, scepticism is moving in the opposite direction. Engagement metrics orient organisational attention toward conversion, leaving fairness implications unexamined. Personalisation vendors supply models trained on data the firm never audits. Campaign workflows rarely include structured review gates. Marketers face career incentives to ship campaigns, not to delay them with internal questioning. Consumer scepticism is informed and rising; organisational scepticism is informed and falling.

The asymmetry has direct consequences for brand legitimacy. Frameworks that rely on informed consumer agency without addressing the asymmetry shift the burden of governance onto the consumers least equipped to carry it. Firms that project confidence in their AI systems while internally lacking the governance to substantiate that confidence face a credibility gap that deepens with every opaque deployment.

This logic produces a counterintuitive prescription. Visible, credible organisational scepticism toward a firm's own AI systems may be a more durable source of consumer trust than projected confidence. The persuasion literature has long shown that admitting limitations signals integrity precisely because admission is costly, and audiences reward the credibility that follows (Crowley & Hoyer, 1994; Eisend, 2006). An organisation that publicly subjects its AI to scrutiny, acknowledges limitations, and shows the seams of its governance signals that it takes its own architecture seriously. For marketing-facing AI, scepticism is both a governance input and a legitimacy signal. Whether this asymmetry holds across industries and consumer segments is an empirical question worth testing.

Takeaway for managers: running a scepticism trace

Designing doubt into governance is not free. Review gates slow timelines. Escalation pathways create organisational discomfort. The evidence suggests that these costs are a condition of performance rather than a tax on it. PwC's 2026 AI Performance study finds that 74% of AI's economic value is captured by 20% of organisations, and the distinguishing feature of those leaders is that they go further on governance: they are 1.7 times more likely to operate a responsible AI framework and 1.5 times more likely to maintain a cross-functional governance board. Their employees are twice as likely to trust AI outputs. The firms winning the AI race are those that have institutionalised doubt.

The scepticism trace is a short diagnostic any manager can run. Take a single consequential AI system in your organisation. Trace one specific doubt through your governance architecture and ask three questions, each tied to a property of scepticism.

Direction. What question is your governance process asking about this system, and is it the question that would catch the failure you are most exposed to? If the primary concern is regulatory compliance, the doubt may be aimed at a narrow target. Ask whether the system's outputs are substantively fair to the people it affects, whether the question is being asked at all, and where in the process it is asked.

Location. Who holds doubt about this system, and where do they sit in the organisation? If the most sceptical actors are organisationally distant from the deployment decision, identify whether a pathway exists for their concerns to reach someone with authority to act. Ask also whether the system is in your governance inventory in the first place. Shadow AI cannot be governed by any architecture that does not see it.

Design. Does your architecture include structured mechanisms that trigger doubt at decision points and, for agentic systems, during operation? Ask whether you can point to a single instance in which internal doubt actually changed an AI deployment decision in the past year. If you cannot, your architecture is producing documentation without accountability.

The firms that will navigate the next decade are those that treat doubt as an organisational resource, designed into governance with the same rigour applied to any other input. For marketing-facing AI in particular, the firms that will retain consumer legitimacy are those willing to be publicly sceptical of their own systems. Treated this way, scepticism is a governance capability, not a posture: something a firm builds into roles, review gates, and runtime monitoring, and then measures by whether internal doubt has changed a deployment decision in the past year. Sceptical, by design.

References

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623).

Blasco-Arcas, L., & Lee, H.-H. M. (2026). From principles to practices: Implementing AI responsibly for social and sustainable impact. In “Encyclopedia of Artificial Intelligence in Marketing”. Cham: Springer Nature Switzerland.

Crowley, A. E., & Hoyer, W. D. (1994). An integrative framework for understanding two-sided persuasion. Journal of Consumer Research, 20(4), 561–574.

Eisend, M. (2006). Two-sided advertising: A meta-analysis. International Journal of Research in Marketing, 23(2), 187–198.

Gartner. (2025, May 15). Emerging Tech: Avoid Agentic AI Failure: Build Success Using Right Use Cases [Research report]. Gartner Insights.

Gartner. (2026). Consumer preferences regarding generative AI in brand communications. Gartner Research.

Grote, G., Parker, S. K., & Crowston, K. (2024). Taming artificial intelligence: A theory of control-accountability alignment among AI developers and users. Academy of Management Review, advance online publication.

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. Minds and Machines, 30(1), 99–120.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389–399.

Klaviyo. (2026). 2026 AI consumer trends report. Klaviyo.

KPMG. (2025). Trust, attitudes and use of artificial intelligence: A global study 2025. KPMG, in collaboration with the University of Melbourne.

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. Nature Machine Intelligence, 1(11), 501–507.

Moffatt v. Air Canada. (2024). 2024 BCCRT 149. British Columbia Civil Resolution Tribunal.

PwC. (2026). 2026 AI performance study. PwC.

Sajadieh, S., Fattorini, L., Perrault, R., et al. (2026). The AI Index 2026 annual report. AI Index Steering Committee, Institute for Human-Centered AI, Stanford University.

Seele, P., & Schultz, M. D. (2022). From greenwashing to machinewashing: A model and future directions derived from reasoning by analogy. Journal of Business Ethics, 178(4), 1063–1089.

Trend Micro. (2026, March 26). Organisations overlook AI risk as governance fails to keep up. Trend Micro.

Tully, S. M., Longoni, C., & Appel, G. (2025). Lower artificial intelligence literacy predicts greater AI receptivity. Journal of Marketing, 89(5), 1–20.

Vakkuri, V., Kemell, K. K., Jantunen, M., Halme, E., & Abrahamsson, P. (2021). ECCOLA: A method for implementing ethically aligned AI systems. Journal of Systems and Software, 182, 111067.