HCRAI

Human-Centric. Responsible AI

THE HUMAN LAYER

Behavioural Risk in AI Systems

The organisations deepest into AI adoption often have the strongest AI governance on paper and the weakest operational controls in practice.

FINDINGS FROM

Wave 1 - June 2026

AUTHOR

Sara Portell

EXECUTIVE SUMMARY

Governance on paper, exposure in practice

The organisations deepest into AI adoption often have the strongest AI governance on paper and the weakest operational controls in practice. That is the central finding of HCRAI's first Behavioural Risk Assessment study, based on responses from **314 AI practitioners** describing real AI systems in their organisations.

FOUR FINDINGS STAND OUT

FINDING 01

43 points

The gap between governance maturity (74 / 100) and operational controls (31 / 100) in organisations with established AI adoption.[1]

FINDING 02

Opposite directions

At each successive stage of adoption maturity, governance grows stronger and operational safeguards grow weaker. Confidence is building faster than the safeguards meant to back it up.

FINDING 03

Invisible risk

Behavioural risk does not appear in dashboards or audit trails, and governance maturity does not reduce it. Exemplary governance reports the same risk profile as almost none.

FINDING 04

1 in 3

AI systems sit in the two highest risk bands, High or Critical. Half are already live, and five in six are live or in real-world pilot.

1 Respondents categorised their organisation's AI adoption as Exploring, Developing, Scaling, or Established, with Established meaning AI is embedded across multiple products or functions, not confined to pilots or a single team.

Underneath these findings sits a more fundamental problem. An organisation can have policies, accountability structures, audit trails, incident and escalation processes, and still not see its users over-relying on outputs, misreading confidence, or how are being shaped or harmed by repeated interaction.

Incident processes detect events, and behavioural risk rarely presents as an event. It accumulates, and by the time it surfaces as a reportable incident, it is usually under a model error or a performance issue. The record captures the symptom, but the behavioural cause stays invisible.

> Governance frameworks describe what organisations *intend*. Behaviour reveals how systems *shape people*.

This study suggests the two have become disconnected, and that the disconnect is wider where AI adoption runs deeper.

314

AI practitioners, each describing one real system

15+

countries across the AI economy represented

4

dimensions: context, behaviour, controls, governance

The neglected dimension of AI risk

AI risk management is maturing globally. The foundational ethical baselines established by the OECD AI Principles[2] have catalysed a wave of regulatory frameworks. The European Union AI Act[3] has the most developed binding requirements, while the United States relies on the NIST AI Risk Management Framework[4] alongside localised state laws. China has enforced iterative vertical regulations, such as its landmark filing and security assessment registry for generative systems[5], while Singapore[6] has pioneered its own Model AI Governance Framework. Across the Global South, emerging economies like India are establishing alternative baselines that balance systemic risk mitigation with digital public infrastructure growth and data sovereignty.[7]

On the enterprise level, multinationals are unifying these requirements by certifying under the international ISO/IEC 42001 standard.[8] These frameworks differ in legal force and geopolitical motivation, but they converge on what they ask organisations to document (e.g., risk classification, technical safeguards and accountability structures). **But convergence on documentation is precisely the limitation this paper is about.**

> *Governance operates at the level of the organisation. Risk materialises at the level of the interaction.*

Risk emerges when a clinician stops questioning a diagnostic suggestion, when a user cannot tell they are talking to a machine, when a hiring tool produces different outcomes for different groups, or when repeated use of an agent erodes the judgment it was meant to support. None of these risks is visible in a policy document. All of them are visible in behaviour.

This interaction layer is a neglected dimension of AI risk. Behavioural risk accumulates gradually, through use, over time. It does not announce itself as an incident. By the time it surfaces, it has usually already shaped decisions, outcomes and people.

2 Organisation for Economic Co-operation and Development. (2024). *Recommendation of the Council on Artificial Intelligence* (OECD/LEGAL/0449). OECD Legal Instruments. legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

3 *Regulation (EU) 2024/1689 (Artificial Intelligence Act).* (2024). Official Journal of the European Union. eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

4 National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). U.S. Department of Commerce. doi.org/10.6028/NIST.AI.100-1

5 Cyberspace Administration of China. (2023). *Interim measures for the management of generative artificial intelligence services.* State Internet Information Office.

6 Personal Data Protection Commission. (2020). *Model artificial intelligence governance framework* (2nd ed.). Singapore.

7 Ministry of Electronics and Information Technology. (2025). *India AI Governance Guidelines.* Government of India.

8 International Organization for Standardization & International Electrotechnical Commission. (2023). *Information technology, artificial intelligence, management system* (ISO/IEC Standard No. 42001:2023). iso.org/standard/42001

THE INSTRUMENT

The four dimensions of the behavioural AI risk assessment

The Behavioural AI Risk Assessment was built to make this layer visible. A structured way for teams building and deploying AI to see their behavioural risk, discuss it in shared terms, and align on what to do. It assesses an AI system across four dimensions. They are not weighted equally. Context and behaviour pull the risk score up; controls and governance pull it down.

Context

PULLS UP ↑

The sensitivity of the use case and audience. Maps to the EU AI Act's risk-based classification.

Behaviour

PULLS UP ↑

What the system is likely to do to the people who use it. **It has no equivalent in existing framework.**

Controls

PULLS DOWN ↓

Technical and operational safeguards built into the product. Correspond to NIST and ISO/IEC 42001.

Governance

PULLS DOWN ↓

Accountability, documentation and oversight structures common to all of these frameworks.

A system used by vulnerable adults with no human review carries more risk than the same system with strong testing, monitoring and clear escalation paths. The overall score reflects this balance. Context and behaviour pull it up, controls and governance pull it down.

THE STUDY

Inside the assessment

This white paper reports findings from **Wave 1**, the first large-scale dataset collected with the instrument. Three of its four dimensions map onto established frameworks, so that what it measures can be understood within existing governance vocabularies.

CONTEXT	Maps to the EU AI Act's risk-based classification, which determines obligations by use case and audience rather than technical performance alone.
CONTROLS	Correspond to the technical and operational safeguards described in the NIST AI RMF and ISO/IEC 42001's operational requirements.
GOVERNANCE	Corresponds to the accountability, documentation and oversight structures common to all of these frameworks.
BEHAVIOUR	Has no equivalent. It is what AI does to the people who use it, not how the organisation manages the AI. It is the dimension this instrument was built to make visible.[9]

In late May and early June 2026, **314 AI practitioners** completed the assessment, each evaluating one real AI system they work with directly. Participants were recruited primarily through the Prolific research platform and screened for professional involvement in the design, development and management of AI systems.

9 The behavioural dimension of AI risk, including cognitive offloading, over-reliance and miscalibrated trust, has been examined in behavioural science research, notably the Behavioural Insights Team's 2025 report *AI & Human Behaviour: Augment, Adopt, Align, Adapt*. The Behavioural AI Risk Assessment operationalises these concerns into a structured risk measure.

THE SAMPLE

Where the data comes from

53% of respondents work in organisations with more than 1,000 employees; a quarter work in organisations above 10,000.

67% are private companies and 22% publicly listed; public sector bodies, academic institutions and non-profits make up the remainder.

59% of organisations are scaling AI in production or have it established across multiple products and functions. Only 8% are still at the exploration stage.

44% of systems assessed are live and deployed to users; a further 37% are in testing or pilot. More than a quarter affect over 5,000 people each.

56% work in technology, with finance, professional services, public sector, healthcare, retail, education and insurance also represented across engineering, data science, product, operations, leadership and compliance roles.

Three quarters act as *deployers*, integrating or configuring AI systems built by others; a third develop systems themselves. Respondents span more than fifteen countries, led by the United States, United Kingdom, Germany, Canada and Spain. The sample is weighted toward North America and Western Europe; coverage of Asia-Pacific, Latin America and the Global South will be a priority for Wave 2.

ACKNOWLEDGEMENTS

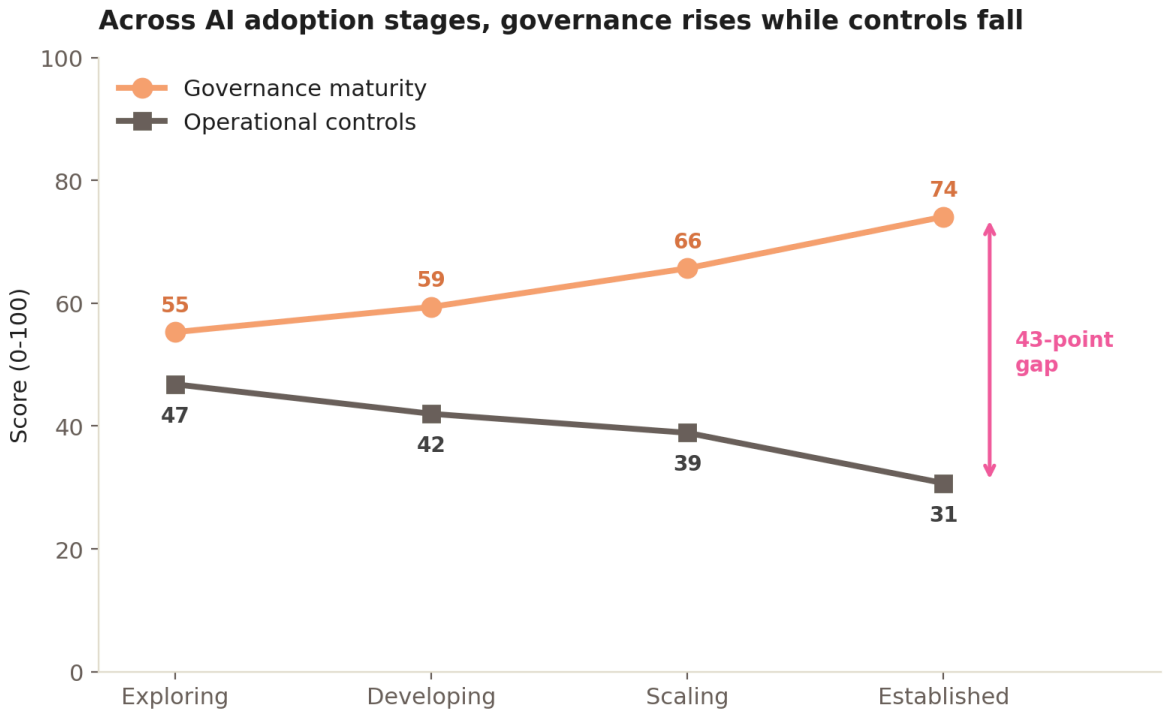
Thanks to Arif Rahman, Danielle Hopkins, Fendi Tsim, Japmandeep Ahluwalia, Silvia Rocha and Yasmina El Fassi, and to early readers, whose feedback and detailed revisions refined the methodology and strengthened the analysis presented here.

FINDINGS

Governance and controls are moving in opposite directions

Organisations still exploring AI report controls and governance at roughly similar levels, a gap of around 8 points. Among organisations where AI is established across multiple products and functions, that gap has grown to **43 points**: governance maturity at 74, operational controls at 31. Both patterns are statistically significant, and they hold independent of respondent role, organisation size and deployment status.

FIGURE 01 n = 314



The dominant pattern across the sample is strong governance sitting on top of weak controls. In organisations at more advanced adoption stages, this misalignment becomes the norm.

WHY THIS HAPPENS

Two mechanisms fit the evidence

One interpretation is reassuring, mature organisations relax product-level constraints because governance compensates. However, the data refutes the premise. Behavioural risk does not fall as governance rises. This compensation is not happening.

01 Structural drift

Governance and controls sit in different parts of the business, and they don't move at the same speed. Governance produces durable artefacts (e.g., policies, committees, registers and audit trails) that persist once created. Controls are product and engineering work that competes with feature delivery every release cycle, and they depreciate. Controls compete with feature delivery and require ongoing attention. They can be weakened, bypassed, outdated over time or fall behind the reality of how the system is being used.

02 A measurement blind spot

Governance functions reasonably believe risk is managed because everything their instruments can see, (e.g., policies, registers and approval records) looks healthy. Nothing in those instruments measures product-level safeguards and user behaviour. The confidence is built on indicators that cannot see the layer where risk materialises. Nearly a quarter of AI practitioners surveyed acknowledged their understanding of how users interact with their AI rests on limited evidence, early assumptions, or no assessment at all.

Whatever the mechanism, the outward result is difficult to distinguish from performative compliance. An organisation with mature governance and eroded controls presents the same risk profile as one that built governance for show. Both carry the same exposure.

This pattern is consistent with emerging independent research. A 2026 study of European professionals and leaders actively using generative AI (n=70), by Zampirolo and colleagues, describes an *AI governance paradox* of ethical awareness without operational clarity, in which organisations recognise AI risks but lack the operational mechanisms to act on them.[10]

The present data extends that picture. The gap is not confined to organisations with immature governance. It is wider where governance is most developed. The **organisations with the strongest governance structures report the weakest operational controls.**

> Behavioural risk is not a future problem. It is accumulating in every interaction happening *today*.

TWO CAVEATS

These are cross-sectional data, and measures are self-reported.[11] Either way, the governance layer is pulling ahead of the operational layer, and the gap between them is growing.

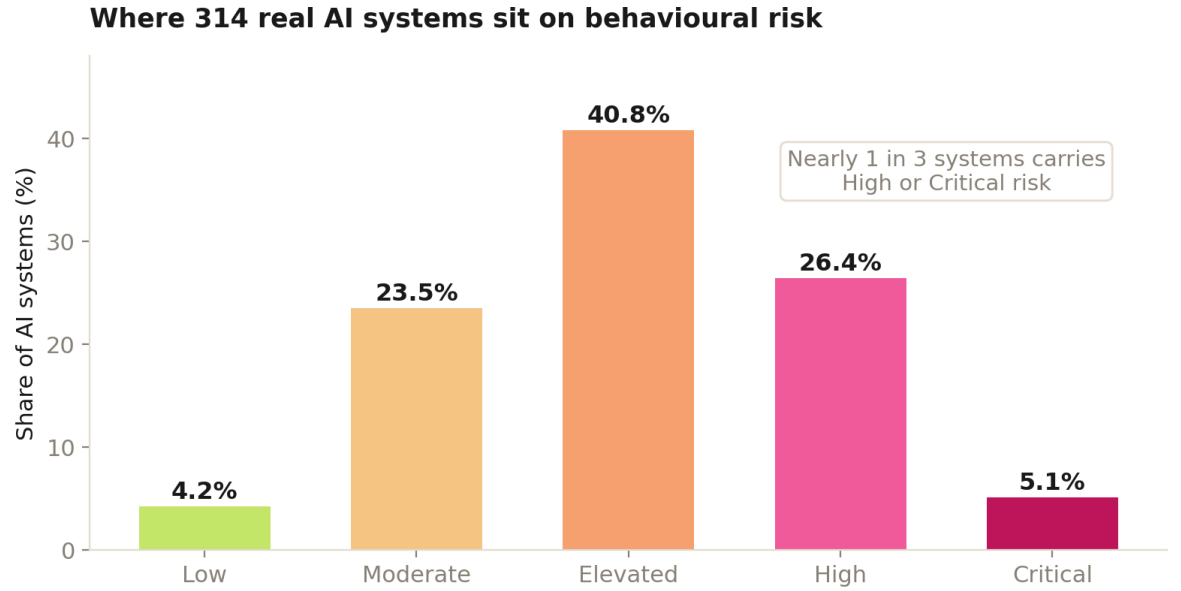
10 Zampirolo, G., Burtscher, F., Mariotti, Y., Rudnik, D., & Aderemi-Makinde, M. (2026). *The AI Governance Paradox: Ethical Awareness Without Operational Clarity.* Artificial Intelligence Ethics, Law, and Policy. IntechOpen. DOI: 10.5772/intechopen.1015694.

11 Cross-sectional data compares organisations at different stages rather than measuring change within organisations over time; Wave 2 will test whether the same divergence appears longitudinally. Because measures are self-reports, part of what differs may reflect not controls themselves but confidence in them, as practitioners in mature organisations become more aware of what is missing.

Behavioural risk across the field

The assessment places each system in one of five risk bands, calibrated against the distribution of the full sample. Relative to the field, nearly **one in three** systems sits in the two highest bands, High or Critical. The single largest group, four in ten, sits at Elevated. Only 4% fall in the lowest band.

FIGURE 02 n = 314



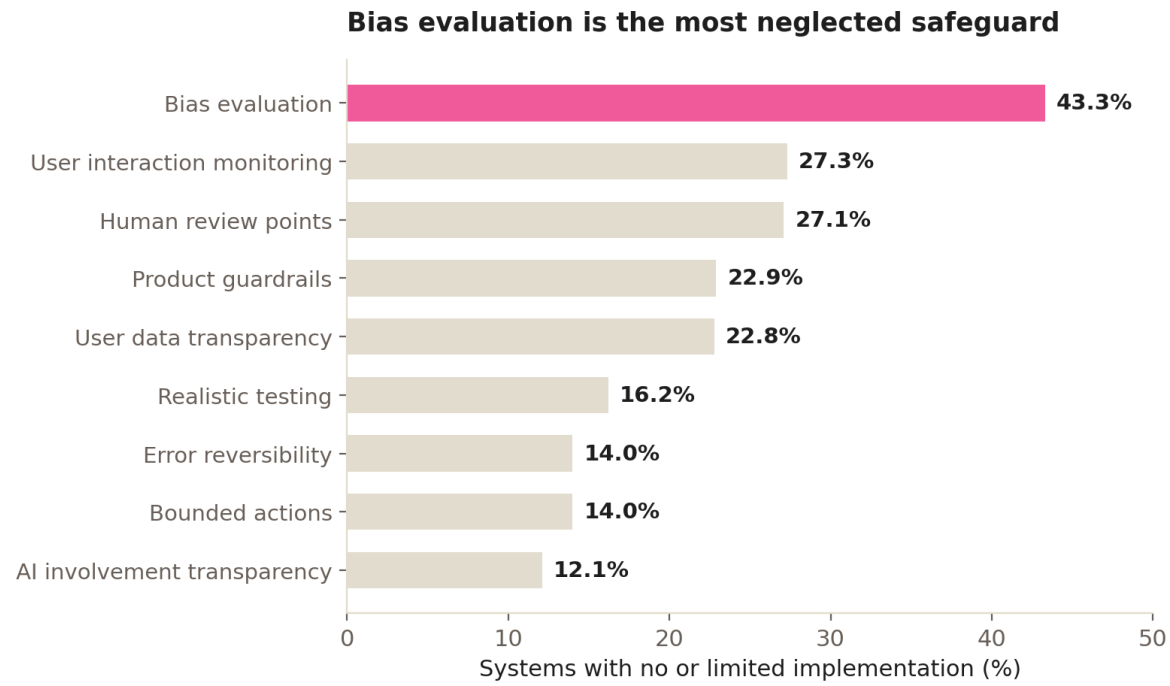
These are not hypothetical systems. Most are live or in pilot, with real users. They include systems used with the general public, adults in sensitive or vulnerable situations and minors, and in domains such as healthcare, finance, education and public services.

WHERE THE GAPS CONCENTRATE

The most neglected safeguard

Bias evaluation stands apart. **43%** of systems have never been evaluated, or only minimally evaluated, for whether they produce different outcomes across different groups of users. That is more than one and a half times the neglect rate of any other control, and the safeguard most directly tied to fairness and regulatory exposure under the EU AI Act.

FIGURE 03

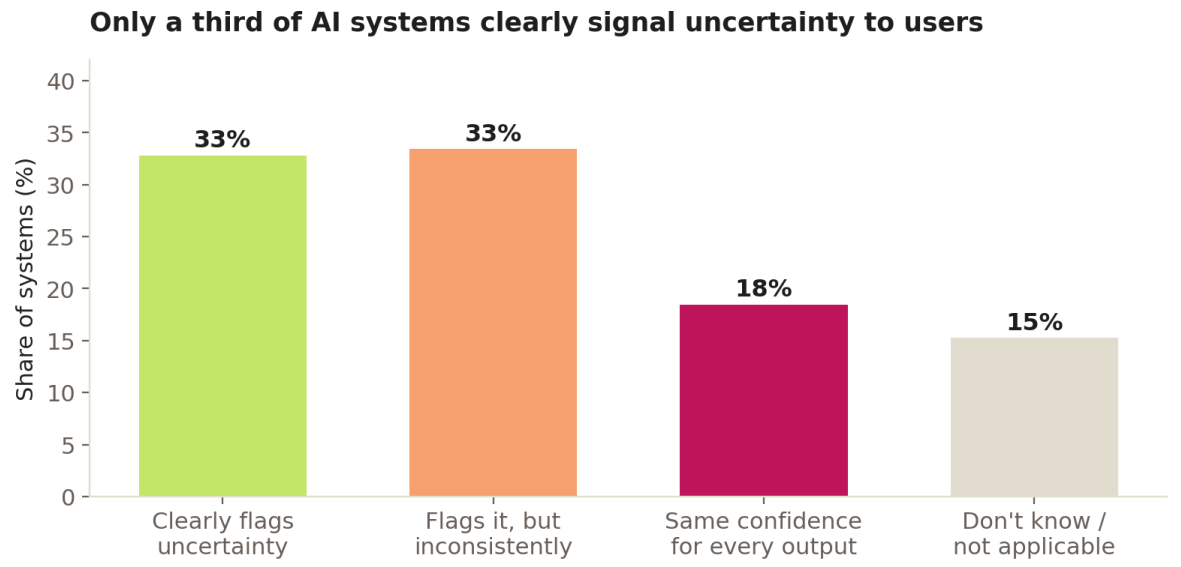


The next most neglected controls are the two that watch the human side of the system. Monitoring how users interact with the AI (27%) and defined points of human review and intervention (27%). **The controls organisations skip are the ones that would surface behavioural risk.**

The confidence problem

One question asked what happens when an output is low-confidence or outside the system's reliable range. Only a third of systems clearly signal uncertainty to their users. Nearly one in five presents every output with the same confidence, giving users no way to calibrate trust. A further 15% of practitioners do not know how their own system behaves in this respect.

FIGURE 04 n = 314



For systems that people rely on to make decisions, this is not a presentation detail. It is the difference between **assisting judgment** and **misplacing certainty**.

Governance maturity does not reduce behavioural risk

Across 314 systems, the correlation between governance strength and behavioural risk is small and not statistically significant.[12] If anything, the relationship trends slightly positive, the opposite of what one would expect if governance reduced behavioural risk.

If stronger governance produced safer user outcomes, respondents in well-governed organisations would still report lower reliance risk, better uncertainty signalling and less harm exposure. They do not, which means the independence is empirical, not an artefact of the instrument. The tools of governance, from policies and accountability structures to audit trails and escalation processes, were simply not designed to see the human side.

~1/4

of practitioners say their understanding of how users interact with their system rests on limited evidence or early assumption; 6% had not assessed it at all.

Controls do work

Systems with stronger product-level safeguards exhibit measurably lower behavioural risk. The layer organisations are neglecting is the one that works.

Organisations are making deployment decisions about systems whose human effects they have not examined. If governance worked the way organisations assume, stronger governance would translate into lower risk in the layer where interaction occurs. The data shows it does not.

12 $r = .105$, $p = .065$; the small positive coefficient is not statistically significant and runs counter to the assumption that stronger governance reduces behavioural risk.

WHAT TO DO NOW

Five things to do now

This pattern is no single function's responsibility. Behavioural risk falls in the space between legal, product, engineering and leadership views, which is part of why it stays invisible. Closing the gap starts with assessing the same system together, not separately.

01 **Audit the gap.**
Most organisations assess governance and product safeguards separately, which is how a 43-point gap stays invisible. Assess them together, per system, and treat divergence itself as a risk indicator.

02 **Make controls mandatory when scaling.**
Controls are weakest where AI is most widespread. Scaling reviews should require evidence that safeguards scaled with the system, not just that a policy covers it.

03 **Start with bias evaluation.**
It is the most neglected control and the most exposed to regulatory risk. Structured evaluation across user groups is a known discipline. There is no defensible reason for 43% neglect.

04 **Make user behaviour visible.**
Monitor how users interact with AI and its effects, not just what it outputs. Reliance patterns and override rates are measurable through product analytics; agency, harm and trust calibration over time require user research, evaluations and red teaming. Most teams are not doing this systematically.

05 **Make uncertainty visible to users.**
Confidence signalling is one of the cheapest behavioural safeguards available, and two thirds of systems do not do it consistently.

ABOUT THIS RESEARCH

A free, structured way to see the human layer

The Behavioural AI Risk Assessment is HCRAI's structured instrument for assessing behavioural risk across four dimensions: context, behaviour, controls and governance. It is available as a free self-assessment at **hcrai.com**, where organisations can profile a system in around ten minutes and see their risk profile.

Wave 1 data was collected in May and June 2026 from 314 AI practitioners, each evaluating one AI system they work with professionally. Wave 2 will recruit a new cohort, allowing comparison over time. The assessment has been submitted to the OECD AI Policy Observatory's catalogue of tools for trustworthy AI.

HCRAI (Human-Centric Responsible AI) applies behavioural science and AI ethics to the design and governance of AI systems, working with organisations to ensure systems are responsible, human-centred, and designed with the people who use them in mind.

GET IN TOUCH

To discuss these findings, assess your own AI systems, or explore what behavioural risk looks like in your organisation:

saraportell@hcrai.com
hcrai.com

THE HUMAN LAYER