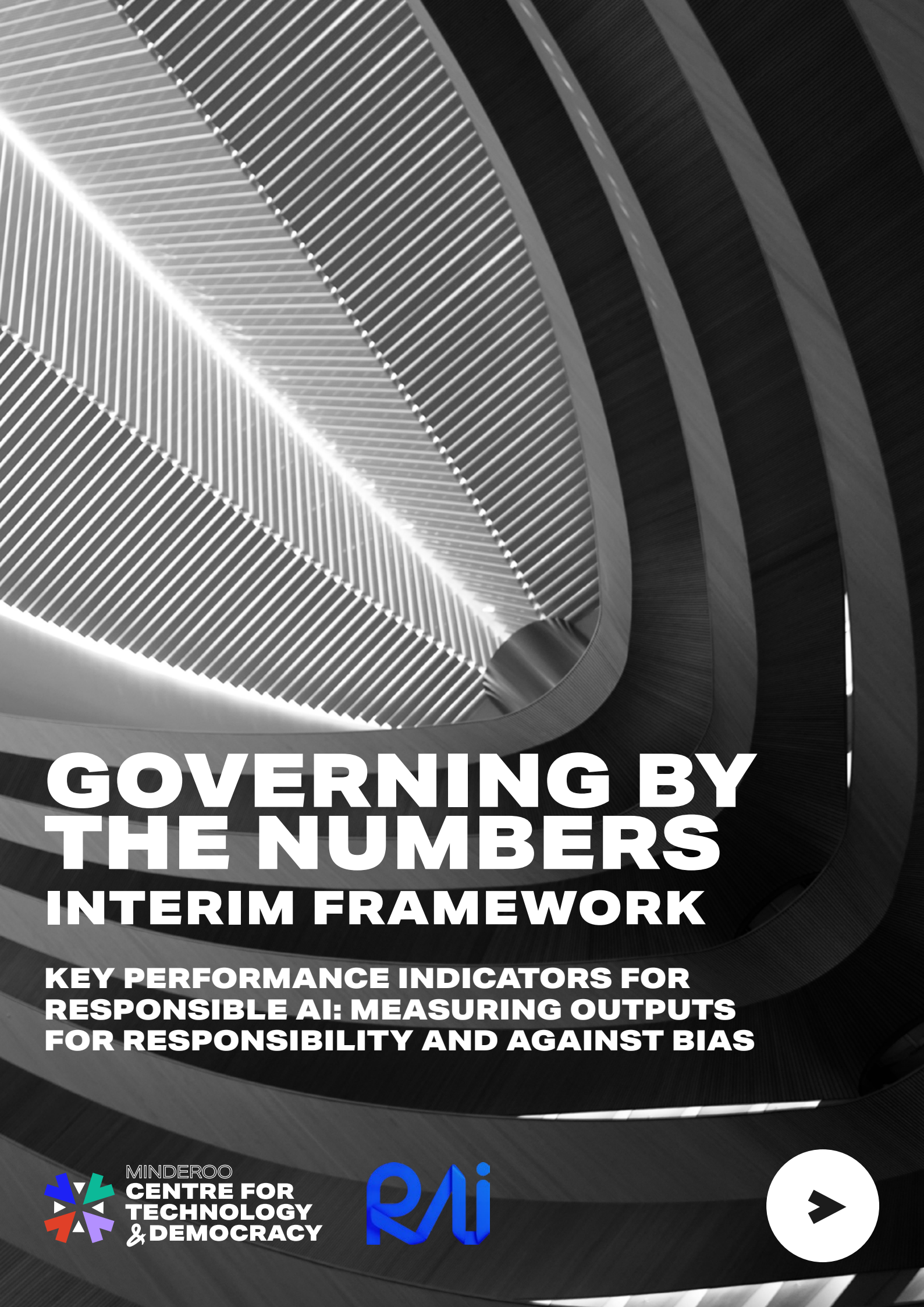


GOVERNING BY THE NUMBERS

INTERIM FRAMEWORK

KEY PERFORMANCE INDICATORS FOR RESPONSIBLE AI: MEASURING OUTPUTS FOR RESPONSIBILITY AND AGAINST BIAS

MINDEROO CENTRE FOR TECHNOLOGY & DEMOCRACY

RAi

This framework is authored by:

Dr James Ward, Research Affiliate,
Minderoo Centre for Technology and Democracy,
University of Cambridge

September 2026

Minderoo Centre for Technology and Democracy

The Minderoo Centre for Technology and Democracy is an independent team of academic researchers at the University of Cambridge, who are radically rethinking the power relationships between digital technologies, society and our planet.

www.mctd.ac.uk

Responsible AI UK

Responsible AI UK (RAi UK) is a UKRI-funded national programme that brings together researchers, industry, policymakers and the public to understand how AI can be developed and used responsibly, for the benefit of people, society and the economy.

www.rai.ac.uk

CONTENTS

INTRODUCTION AND SUMMARY 4

1. WHAT ARE AI GOVERNANCE METRICS TODAY? 5

Where in the AI value chain should we measure? 5

EU and US Posture 6

The UK's Approach 6

The cost of failure, and a necessary nuance 7

2. THE HARMS OF NOT ACTING NOW 8

Apple Card (2019): the disparity nobody was required to compute 8

COMPAS (2016): duelling metrics with no referee 8

Facial recognition wrongful arrests (2018–2025): known error rates, unmonitored use 9

3. A KPI FRAMEWORK FOR RESPONSIBLE AI 9

AI KPI Framework 10

4. FROM MEASUREMENT TO TRIGGER: THE SRL MATRIX 12

The mechanics 12

Accountability for claims, not explanations 13

5. RECOMMENDATIONS FOR THE UK 14

For policymakers and regulators 14

6. CONCLUSION 15

INTRODUCTION AND SUMMARY

We urgently need to know if AI does what it says on the box, especially as public services reach for AI tools for a wide variety of tasks.

It feels simplistic, but measuring whether a tool provides the benefits it was deployed to provide is basic management and basic governance. What we see more frequently are governments and enterprises alike using the success measurements the AI companies themselves offer: How large is the context window? How many tokens can be processed in a given timespan? These answers are helpful, but they presuppose the more important, earlier questions: Does using an AI tool for [x] purpose make sense? What benefit do we expect to get beyond being able to say "we're using AI?" In other words, does the AI do what it says on the box?

To determine whether using AI would deliver value and usefulness, we first need to decide what value and usefulness mean. Not taking that step carries risk. Resources wasted, trust broken, communities harmed — the consequences of rushing headlong into AI deployment are well documented. What has not emerged, yet, is a straightforward method for scoping the risk/reward ratio for AI use by governments and the public sector.

To address this gap, MCTD will publish a longer review of performance metrics, we call them KPIs or *key performance indicators*, used by governments in the last 50 years.

This interim paper presents **a structured KPI framework and a practical classification tool, the Cumulative Size and Risk Level (SRL) matrix,** which links an AI system's reach and criticality to escalating, enforceable oversight obligations. Metrics therefore become not just measurement tools, but governance triggers.

Our framework is flexible, practical, and data-driven. With this work we do not try to solve every problem in AI deployment. Rather our aim is to provide decisionmakers with a shorthand for identifying risk in the AI systems that they use and the language to explain it. Most importantly, we provide a structure that is understandable without requiring specialised technical knowledge.

Our central message is simple: what gets measured gets managed. The UK, with its deep institutional experience of performance governance and its stated ambition to lead on AI assurance, is uniquely placed to become the first jurisdiction where responsible AI can be demonstrated by the numbers.

1. WHAT ARE AI GOVERNANCE METRICS TODAY?

AI governance is not a blank slate. Across jurisdictions, governments and institutions have articulated high-level ideals for the development and use of artificial intelligence, but they have offered relatively little in the way of concrete measures to evaluate whether those ideals are being met in practice. Prevailing approaches, from ethical guidelines to emerging regulation, remain largely aspirational: they outline what good AI governance should look like, but stop well short of specifying how to quantify success or failure in achieving it.

The recent proliferation of AI governance frameworks follows a now-familiar pattern: an escalating volume of ethical principles, risk taxonomies and procedural obligations, paired with a conspicuous absence of performance standards or outcome metrics. The result is a metrics gap: a disconnect between the broad, principle-based language of AI governance and the performance-driven tools needed to hold actors accountable. To date, no jurisdiction with major regulatory power has articulated a framework of governance metrics that it is possible to fail to achieve.

In the absence of clear benchmarks or key performance indicators, **it is difficult to tell whether an AI governance regime is genuinely protecting the public and upholding values, or merely providing comforting rhetoric.** This gap undermines accountability, and it also blocks learning and improvement: without measurement, how do policymakers know what is working and what is not?

Where in the AI value chain should we measure?

'AI failures' are frequently compound failures: responsible-AI metrics must be situated within the broader delivery pipeline, rather than treating 'the AI' as a single object of measurement. An algorithm may appear unfair when the true failure was upstream: a developer never collected data in the right way, so what surfaces downstream as an 'AI bias issue' is at root a data management failure.

Equally, failures can occur downstream of a perfectly sound model: a facial recognition system may be technically reliable and yet be misappropriated by a police force, deployed outside its validated conditions, or so poorly communicated that users misunderstand its limits.

A credible measurement regime therefore assigns indicators to distinct stages of the pipeline:

- **Data practices:** representativeness, provenance, documentation and quality of training and evaluation data (where many 'bias' problems actually originate);
- **Model behaviour:** accuracy, error rates and their distribution across groups, robustness, and performance under distribution shift;
- **Deployment and use:** whether the system is used within its validated scope, with appropriate human oversight, escalation routes and controls;
- **Communication and redress:** whether affected people are informed, can understand and can contest outcomes.

Locating each KPI at the correct stage prevents characteristic errors of AI accountability debates: blaming 'the algorithm' for organisational failures and allowing organisations to hide behind 'the algorithm' when their own data, deployment or communication practices are at fault.

EU and US Posture

The EU AI Act is the world's first comprehensive AI law. Compliance under the Act is evaluated through conformity to process, not through system outputs. The EU AI Act requires providers to assess and mitigate risks but does not measure whether those steps result in safer or fairer AI in actual practice. The effect, consequently, has been limited — despite the fact that the high-risk obligations slipped from August 2026 to December 2027, no one can say whether the delay cost anyone anything, because there are no outcome measures to be behind in reaching.

The US, on the other hand, has shrugged off regulation almost altogether at the federal level. The Blueprint for an AI Bill of Rights was avowedly non-binding, with no quantification. Congressional legislative proposals continue to circulate but as of this writing there is no unified US metrics regime, and any administrative metrics that emerge will be post hoc, piecemeal and jurisdiction-specific.

The UK's Approach

The UK has taken a deliberately light-touch, principles-based path since the 2023 white paper 'A Pro-innovation approach to AI regulation' set out five cross-sector principles (safety, transparency, fairness, accountability, contestability) for existing regulators to apply. DSIT's October 2025 'Blueprint for AI regulation' confirmed the preference for principles-based, sector-led regulation over horizontal statute, with the AI Growth Lab programme of regulatory sandboxes as its centrepiece as of last June, and the Regulating for Growth Bill announced in the May 2026 King's Speech would provide cross-economy sandboxing powers.

The UK has developed Algorithmic Transparency Recording Standard (ATRS), a standardised, two-tier public record of how and why public sector organisations use algorithmic tools. It has been mandatory for all government departments, and for arm's-length bodies delivering public or frontline services, since the phased rollout that began in 2024; well over a hundred records are now published on GOV.UK, the government's 'Blueprint for Modern Digital Government' commits to building on it, and the OECD has identified it as world-leading. The ATRS is a genuine achievement: it proves that mandatory, standardised, cross-government disclosure of algorithmic tools is administratively feasible in the UK.

But the ATRS records what systems are — their purpose, design, data and risks — not how they perform. It is a transparency register, not a performance regime: it contains no outcome indicators, no thresholds, and no consequences that follow from what a record reveals. The same is true of the UK's other assurance instruments: AISI's evaluations, the AI Cyber Security Code of Practice, and voluntary self-assessment tools create sites for evaluation but define no thresholds; they encourage assessment but impose no consequences.

And so, even in the UK, there is no functioning governance regime in which performance metrics act as enforcement triggers, which is precisely what is needed.

The cost of failure, and a necessary nuance

Disparities in credit decisions, error rates in risk prediction, false-match rates in facial recognition: policymakers now are well aware of a litany of failures in AI deployment. In most cases, the risks were quantifiable, and quantification is precisely what eventually enabled accountability, allowing regulators to investigate, courts to scrutinise, and cities to legislate. A proactive, metrics-informed regime would have surfaced each issue before it became a scandal.

In medicine, differential performance across demographic groups is not unique to AI, but it is an accepted and openly managed reality. Diagnostic baselines and disease prevalence differ across populations, data gaps and imbalances are endemic, and clinicians are trained (if imperfectly) to take these known equity issues into account in diagnosis. These challenges are not unique to AI systems.

What AI changes is not the existence of differential performance but its visibility. An algorithm can silently automate and scale a disparity that a clinical culture at least nominally acknowledges and adjusts for. This strengthens rather than weakens the case for measurement. The goal of equity KPIs is to restore, in algorithmic form, the explicit awareness of differential performance that good professional practice already demands, and to distinguish between disparities rooted in data gaps, in model design, and in how a system is deployed and communicated.

2. THE HARMS OF NOT ACTING NOW

To make the stakes concrete, consider three high-profile AI controversies from the US. In each, no proactive metric or reporting requirement existed that might have flagged the issue early; problems came to light only through social media outrage, investigative journalism or the courts — after the harm was done. And in each, the harm was quantifiable all along.

Apple Card (2019): the disparity nobody was required to compute

When customers noticed that women appeared to receive far lower credit limits than men with similar or stronger financial profiles — including, famously, the spouses of prominent technology figures — the New York State Department of Financial Services opened an investigation into the card's underwriting. The investigation, concluded in 2021, did not find unlawful discrimination under fair lending law, but it sharply criticised the opacity of the credit decisioning and the deficiencies in customer service and appeal that turned individual anomalies into a public crisis.

The instructive point is what was absent: no standard metric or required bias audit obliged anyone to monitor approval rates or credit limits by demographic group before launch. A simple KPI — approval rate and limit distribution by gender within credit-score bands — would either have surfaced a genuine problem early or given the firms the evidence to rebut the allegation immediately. Instead, the absence of measurement left everyone, including the companies, arguing in the dark.

COMPAS (2016): duelling metrics with no referee

ProPublica's analysis of the COMPAS recidivism tool used in US courts found that Black defendants who did not reoffend were nearly twice as likely as white defendants to have been classified as high risk — a stark disparity in false positive rates. The vendor responded that the tool was fair on a different metric: predictive parity at each risk score.

Both claims were arithmetically true, but the deeper failure was institutional: no oversight body had specified which fairness metrics counted, required their routine reporting, or audited them. A tool influencing liberty decisions operated for years across jurisdictions with no mandated disaggregated error reporting, and the disparity was discovered by journalists rather than by any accountable institution.

Facial recognition wrongful arrests (2018–2025): known error rates, unmonitored use

A series of documented wrongful arrests in the United States — disproportionately of Black men and women, beginning with Robert Williams in Detroit — followed police reliance on incorrect facial recognition matches. This was foreseeable: NIST's 2019 demographic evaluation of face recognition algorithms found substantially higher false-match rates for some demographic groups across many algorithms, and academic audits had reached similar conclusions.

Yet departments adopted the systems with no operational metrics — no tracking of false match rates by demographic group in live use, no validation on local data before deployment, no audit of how often matches led to arrests later invalidated. Each wrongful arrest was treated as an isolated mistake rather than the statistically expected consequence of a system with known, measurable error properties deployed without monitoring.

3. A KPI FRAMEWORK FOR RESPONSIBLE AI

Below, we present a framework of KPIs for AI governance, aimed at translating high-level principles into operational metrics. The framework is most directly applicable to public sector and government applications, but it contains lessons which are useful for private sector contexts too.

The structured KPI framework spans five categories: Demographic Equity, Transparency, Accountability, Impact, and Adaptability. It includes a practical classification tool, the Cumulative Size and Risk Level (SRL) matrix, which links an AI system's reach and criticality to escalating, enforceable oversight obligations. Metrics become not just measurement tools but governance triggers.

We know that not everything that matters can be perfectly measured, but governance requires that we measure what we can, honestly labelled as proxy rather than essence.

The aim of a proxy metric is not to claim it fully represents a value like fairness, but to give a window into it: a foothold for evaluation, comparison and intervention. One unseen but crucial component of this kind of framework is its channelling function, requiring people to ask critical questions that they might not otherwise have considered. That reframes the project and prompts a very simple corollary question: whom will this benefit and whom might it harm?

Every indicator in this framework is designed to provide raw data (for audit and research), graded scores with plain language anchors (for decision makers and the public), and a trendline over time (to show whether things are improving or not). These frameworks are not an exhaustive list, and provide a starting guide for what measurements might (and should) look like and how such a list of measures could be produced.

AI KPI FRAMEWORK

1. Demographic Equity
1.a Disparate impact ratio. For systems making or informing decisions about people, the ratio of favourable-outcome rates between demographic groups. The four-fifths convention long used by the US Equal Employment Opportunity Commission (flagging ratios below 0.8) provides a familiar starting anchor: not a rigid legal threshold, but a trigger for justification. Operators of high-stakes systems should report these ratios for protected characteristics and either justify disparities as job- or risk-related, or remediate them.
1.b Error-rate parity. A diagnostic tool that is 95 per cent accurate for men and 80 per cent for women is a serious, measurable equity failure — typically rooted in training-data gaps, which is where remediation could be targeted. Policy can require that key performance metrics not differ by more than a stated margin across major groups without a documented mitigation plan.
1. c Bias incident rate. Logged and categorised reports of biased outputs, normalised by decision volume, with the accessibility of the reporting channel itself measured so that suppressing reports is detectable.

2. Transparency
2.a Documentation completeness. A scored rubric tied to required documentation about intended use, training data characteristics, evaluation results, etc. This deliberately builds on what model cards and datasheets already do. The policy contribution is to standardise all required elements, have completeness independently scored rather than self-assessed, and make a minimum score a condition of deployment for high-impact systems.
2.b User awareness rate. Measured through periodic surveys and compliance checks: what proportion of affected people knew an algorithm was involved in a decision about them and knew how to seek explanation or review. Targets approach 100 per cent where disclosure is required.
2.c Explanation provision. The proportion of adverse automated decisions accompanied by an intelligible explanation, and the median time to fulfil explanation requests — rights that UK data protection law (as amended by the Data (Use and Access) Act 2025) increasingly makes concrete for solely automated decisions.
2.d Public registration. The proportion of in-scope systems recorded in a public register. Here the UK is ahead: the Algorithmic Transparency Recording Standard already operationalises exactly this metric for central government. The extension paths are to the broader public sector (where ATRS remains voluntary), to local government, and — for high-impact systems — to the private sector.

3. Accountability
3.a Audit follow-through. Not merely whether audits occur, but time between audits, the proportion of audit findings remediated within the committed period and repeat findings across audit cycles. Accountability means acting on audits, not accumulating them.
3.b Incident response time. Mean time from identification of a significant AI failure, harm or bias finding to mitigation, benchmarked within and across sectors, with severity-graded expectations (critical safety issues in days; others on defined timescales).
3.c Redress accessibility and outcomes. Relying on historical overturn rates alone is risky: a low rate of successful appeals may mean the system is accurate, or that the appeal route is inaccessible — the number cannot distinguish. The metric therefore foregrounds process accessibility metrics: the number of steps required to lodge an appeal; availability of non-digital routes; median time to a human decision; the proportion of affected people who can identify how to appeal (surveyed); and cost to the appellant.
3.d Regulatory compliance. At sector level: enforcement actions issued, proportion complied with, and time to compliance — reported annually to Parliament so that the oversight system's own teeth are visible.

4. Impact (Safety and Societal Outcomes)
4.a AI incident rate. Defined incidents (significant malfunction, harm or near-miss attributable to an AI system) per unit of operation, on the model of aviation and clinical adverse-event reporting, with safe-harbour provisions to encourage candid reporting. Reported incidents will typically rise before they fall as monitoring improves — a pattern that regulators should anticipate and communicate. This is effectively a measurement of how likely harms are to occur.
4.b Public confidence tracking. Regular, methodologically consistent survey measures of public confidence in AI in specific domains (health, policing, finance).
4.c Blindside rate. Of realised AI harms in a period, the proportion arising from systems classified as low-risk or outside the regime's scope. A rising blindside rate is direct evidence that the risk classification needs revision — the metric that keeps the framework honest about its own coverage.

5. Adaptability (Governing the Governance)
5.a Review cadence. Whether scheduled reviews of the indicator set thresholds and risk classifications occur and lead to published revisions. Regular revision is also the structural defence against gaming: a metric that moves is far harder to teach to.
5.b Evidence engagement. Expert input received and acted on: advisory processes held, recommendations adopted, research commissioned — a feedback loop from the research community into the regime, reported publicly.
5.c Horizon-scanning performance. Retrospective self-assessment, published for comment: which emerging risks did the regime identify in advance, and how quickly did guidance follow major technological developments?

4. FROM MEASUREMENT TO TRIGGER: THE SRL MATRIX

The datapoints generated via the framework, on their own, are interesting but not necessarily actionable. To make the data provided meaningful and predictive of future outcomes (such as they can be measured) there must be an evaluation component.

A transparent way to decide which systems owe which obligations is required, so that a village lending library's recommendation engine is not regulated like a national diagnostic platform. Risk-based regulation typically answers this with qualitative categories, such as 'high-risk' or 'critical use', whose interpretive ambiguity generates both over-compliance and evasion. We propose a simple, legible alternative: the Cumulative Size and Risk Level (SRL) matrix.

The mechanics

Each AI system is scored on two dimensions. Its **Size Tier (ST1–ST5)** reflects the population scope it affects, from individual or very limited reach (ST1) through tens of thousands (ST3) to national or global user bases (ST5).

Its **Risk Tier (RT1–RT5)** reflects the severity and reversibility of its decisions, from low-stakes assistive functions such as music recommendation or grammar correction (RT1), through consequential but recoverable domains such as content moderation or routine financial decisions (RT3), to life-altering applications in medical diagnosis, autonomous navigation or criminal justice (RT5).

The sum — the **SRL score, from 2 to 10** — is a governance signal that maps to escalating obligations:

- **SRL 2–3 (minimal oversight):** registration and basic documentation only. Even low-risk systems disclose their existence and function — the ATRS logic — without burdens that would stifle harmless innovation.
- **SRL 4–7 (heightened oversight):** formal attestation of conformity with relevant standards; fuller documentation of design, data and testing; periodic reporting of KPIs appropriate to the system; internal audit on a defined cycle.
- **SRL 8–10 (stringent oversight):** independent third-party evaluation before and during deployment; senior executive attestation of compliance, placing accountability at the top of the organisation; full KPI reporting with regulator access to underlying records; and enforceable consequences — including suspension from specified uses — for failure to meet obligations.

Two properties distinguish this from existing risk-tiering.

- First, the score is dynamic: if a system's user base grows from thousands to millions, or it is repurposed from a low-stakes to a high-stakes context, its rising SRL score automatically moves it into a more demanding bracket — new obligations are triggered by measured reach and use, without waiting for a mishap to prove that more oversight was needed. Metrics thereby operate as gateways and checkpoints, not just retrospective report cards.
- Second, the inputs are auditable facts (user counts, deployment contexts) rather than self-assessed adjectives, narrowing the room for classification games that qualitative labels invite.

Accountability for claims, not explanations

The SRL approach embeds a further lesson for the deep-learning era: effective accountability can rest on verifiable claims and outcomes rather than on fully explainable mechanisms. For complex models, demanding a complete causal account of each decision is frequently futile.

The SRL matrix does not hinge oversight on prying open the black box. Instead, operators must state in attestations and documentation what the system will and will not do: its intended performance, safety measures and limitations.

Oversight can then verify whether real-world behaviour matches the claims, through test results, audits and ongoing KPI reporting. A firm that attests its diagnostic tool meets a stated accuracy and equity profile is held to that profile; significant deviation triggers penalties, mandated improvement or withdrawal.

A company or a government may not be able to explain every decision their AI makes, but they would be responsible for ensuring its measured behaviour meets the standards that they claim for it.

The SRL matrix is offered as an exemplar rather than a finished statute: tier boundaries, obligation menus and sectoral calibrations are exactly the kind of parameters the UK's forthcoming regulatory sandboxes exist to test. Its purpose is to demonstrate that KPIs can serve as the backbone of a responsive governance framework — one in which oversight escalates in a calibrated, predictable, rules-based way as a system's footprint and stakes grow.

5. RECOMMENDATIONS FOR THE UK

The UK enters this agenda with unusual advantages: three decades of performance-governance experience, the world's most developed public-sector algorithmic transparency register, a frontier evaluation institute, regulator-led model risk discipline in financial services, and — in the AI Growth Lab and the sandbox powers of the Regulating for Growth Bill — a legislative vehicle designed for exactly this kind of controlled experimentation.

For policymakers and regulators

1. **Introduce a duty to measure.** Introduce a duty on when high-impact AI systems are procured and used by public sector bodies to measure and report a defined set of responsible-AI KPIs — with independent verification for the highest tiers and consequences attached to defined thresholds. Regulation should require outcome reporting, not only process compliance.
2. **Build on the ATRS.** Extend the ATRS from recording what systems are to reporting how they perform: add a performance tier to ATRS records for in-scope tools (disaggregated accuracy and error rates, incident and redress statistics), extend mandatory scope beyond central government to the broader public sector and local government, and pilot private-sector registration for high-SRL systems. Align ATRS to new and upcoming AI standards and host the ATRS maintenance within a regulatory body that can ensure it is kept up to date. Develop automation tools to standardise reporting and build it into procurement requirements. In the long term, embed evaluation tools alongside ATRS.
3. **Pilot the SRL matrix in the AI Growth Labs.** Use the forthcoming sandbox powers to trial SRL classification and KPI reporting with volunteer firms in two or three sectors (recruitment technology, medical AI, and public-sector decision support are natural candidates), before any wider mandate. Sandboxes exist to de-risk exactly this.
4. **Create the measurement infrastructure.** Develop standard definitions for the core indicators, an accreditation scheme for third-party evaluators, and a safe-harbour incident reporting system modelled on aviation safety reporting.
5. **Report to Parliament.** Lay an annual State of Responsible AI report presenting sector-level KPI aggregates, enforcement statistics and blindside analysis — so that the governance regime itself is accountable by the numbers, and so that Parliament can judge whether pro-innovation and pro-accountability are in fact being delivered together.

6. CONCLUSION

AI governance must move from a principles-only approach to one grounded in measurable accountability. That requires measuring the inputs and outputs that matter to the people who will be affected.

We have proposed a system of KPIs that can be monitored, enforced and refined — situated correctly along the AI value chain, built on the measurement instruments that industry and the UK public sector have already created, reported in forms people can actually interpret, verified independently, and attached to proportionate consequences through mechanisms like the SRL matrix.

Demanding hard data on outcomes will drive organisations to internalise governance goals; using that data to adapt the rules keeps the regime relevant as the technology moves. Governance itself becomes a learning system: setting targets, measuring results, updating strategy — exactly what we expect of the systems under regulation.

Moving from abstract principles to operational KPIs is the maturation of AI governance: a progression from saying 'do the right thing' to rigorously verifying that *the right thing is being done.* It preserves the flexibility to adjust as we learn, while providing the clarity needed to hold every actor accountable.

MINDEROO

CENTRE FOR TECHNOLOGY & DEMOCRACY

The Entopia Building
1 Regent Street, Cambridge
CB2 1GG

www.mctd.ac.uk

minderoo@cisl.cam.ac.uk

UNIVERSITY OF CAMBRIDGE

Design by **CHRIS PRINTED THIS**

License free image taken from Unsplash courtesy of Ilia Bronskiy