Weight sensitivity of leadership in the CAIDP AI Index

Lluis Eriksson

November 2026

Abstract

Canada and Japan share the highest numerical score in both the 2025 and 2026 CAIDP AI Index. How differently must the twelve indicators be weighted before another country can tie them? The smallest possible largest-to-smallest weight ratio falls from 3 to 1.5. Restricting both editions to the same 80 countries gives the same thresholds. In a crossed comparison of the two response tables, replacing just the leaders' evaluation on implementation of UNESCO's AI ethics recommendation accounts for the entire change in the threshold. The other countries' ratings determine which challengers can enter at the boundary. These findings concern the published evaluations; their interpretation as changes in national institutions requires additional evidence.

1 The comparison

An index reader may attach more importance to an enforceable right to algorithmic transparency than to endorsement of international principles. We ask how far weights must depart from equality to change the countries sharing the highest score. Sensitivity to criteria weights is an established subject in multicriteria decision analysis [3]. Our contribution is a checkable boundary for two editions of a specific policy index and a decomposition of its change.

The Center for AI and Digital Policy (CAIDP) evaluates twelve questions. In the public map, yes receives one point, partial receives one half and no receives zero. The sum reconstructs every displayed score in 2025 (80 countries) and 2026 (90). All 80 countries from 2025 remain in the 2026 map. We compared the twelve question headings in the full reports, on physical PDF pages 1457–1462 and 1731–1736 respectively. Their wording agrees after normalization of spacing, case and punctuation. The underlying grading instruction for Q11 did change. In 2025 the methodology recognized leadership in developing and implementing the UNESCO recommendation for full credit and cited completion of the UNESCO Readiness Assessment Methodology (RAM) as one example of implementation. In 2026 full credit explicitly requires both completion of the RAM and significant further policy action. The printed country grids independently confirm the change on Q11 for Canada and Japan. This documents a break in the scoring rubric; it does not establish the reason for either individual rating.

In 2026 we also paired every printed evaluation-grid row to its map record. All 1,080 responses and 90 scores agree. Mauritius has 7 points in both; its tier label is IV in the printed grid and III in the map. Our outcome is the set attaining the largest numerical score. Six countries occupy the highest published tier in 2026, and two attain the maximum score.

2 A sharp boundary

The positive weights may be rescaled so that $1 \leq w_i \leq R$. Canada and Japan have identical response vectors within each edition. Write c for that common vector and x for a challenger. Put

$P_x = \sum_i (x_i - c_i)_+$ and $N_x = \sum_i (c_i - x_i)_+$. Since each coordinate can be weighted separately, the largest possible weighted advantage of x is $RP_x - N_x$. It is attained by giving weight R precisely to the coordinates on which x gains. For a challenger initially below the leaders, its entry threshold is therefore N_x/P_x if $P_x > 0$ and infinity otherwise. The minimum finite threshold across challengers is exact: below it every outsider loses, and at it at least one outsider ties at the top. The duplicated leader vector matters. Different leaders in different directions would require joint comparisons.

Edition	Minimum R	Countries first able to enter
2025	3	Netherlands, South Korea, United Kingdom
2026	3/2	Netherlands

In 2025 Canada and Japan each score 11.5 out of 12. Their only partial answer is Q9, on a legal right to algorithmic transparency. In 2026 both score 11, with partial answers on Q9 and Q11. The latter asks about implementation of UNESCO's Recommendation on the Ethics of AI. The Netherlands in 2026 gains half a point against the leaders on each of Q9 and Q11 and loses half a point on each of Q2, Q10 and Q12. Giving weight 1.5 to Q9 and Q11 and unit weight to the remaining questions yields 11.5 for each of the three. No other country can reach that score under those weights. The threshold is a tie boundary; the Netherlands can strictly overtake the pair if those two weights increase further. Even allowing arbitrarily disparate positive weights, 47 of the 88 outsiders can never match the leaders in 2026.

A different perturbation leaves the leaders unchanged. With equal item weights and common partial credit t, a country's score is $(1-t)y + t(y+p)$, where y, p count its yes and partial responses. In 2026 only the two leaders have ten yes responses, and their total of yes plus partial responses is twelve. They therefore strictly lead every outsider for $0 \leq t < 1$. At $t = 1$ sixteen countries tie. The distinction matters when asking which design choices influence the ranking.

3 Where the change arises

We repeat the comparison among the same 80 countries and cross the leaders' response vector from each edition with the other 78 countries' response vectors from each edition. The mixed tables are artificial comparisons. In particular, they do not represent an additional observation of any country.

Leaders	Others	Minimum R	Boundary entrants (ISO codes)
2025	2025	3	GBR, KOR, NLD
2025	2026	3	CHE, GBR, NLD, NOR
2026	2025	3/2	KOR, NLD
2026	2026	3/2	NLD

The ten additional countries in 2026 do not alter the threshold. Holding either set of challenger ratings fixed, changing the two leaders' Q11 from yes to partial suffices to reduce the boundary from 3 to 1.5. Holding the leaders fixed and changing the other countries' evaluations changes the identities of boundary entrants while preserving the threshold. Thus the index producer's coding and evidence for one item merit particular attention when readers compare the robustness of the headline ranking across editions.

4 Scope, sources and review context

The weight box allows any positive importance ratios up to R. We have no survey evidence that users employ those weights. An evaluation may change because the underlying evidence, rubric or practice changed. The crossed table measures a dependency of the reporting system; it cannot apportion those historical causes. The implementation reads the publisher's public map as inert data, verifies its checksum, computes rational bounds and tests the extremal weights against the entire cohort. The publisher retains rights in its reports and response tables.

Model assessments and review context. OpenAI Codex assisted with the analysis and writing. The author administers AIRR. Any version-specific assessment of the exact PDF is documented separately with its model identity, involvement and objections. No independent peer review is claimed.

References

[1] Center for AI and Digital Policy. *Artificial Intelligence and Democratic Values, CAIDP Index 2025*. `https://www.caidp.org/reports/caidp-index-2025/`.

[2] Center for AI and Digital Policy. *Artificial Intelligence and Democratic Values, CAIDP Index 2026*. `https://www.caidp.org/reports/caidp-index-2026/`. Interactive data at `https://ai-values-map.vercel.app/`.

[3] B. Mareschal. Weight stability intervals in multicriteria decision aid. *European Journal of Operational Research* 33 (1988), 54–64. `https://doi.org/10.1016/0377-2217(88)90254-8`.