Human–AI Collaboration Score Methodology

How much human editorial judgment shapes each published digest

Purpose

The Tech for Human Agency Pulse is a weekly digest of Ashoka Fellow activity in technology and human agency, produced through a structured human–AI editorial workflow. AI generates candidate stories from Fellow social media and public sources; a human editor reviews, filters, edits, and publishes the final digest.

The Human–AI Collaboration Score measures how much human editorial judgment shapes the published output. Its primary function is an **internal quality signal**: flagging when the editorial process risks becoming a rubber stamp rather than a genuine curation exercise.

The 60/40 Model

The total score is 100 points, split between AI and human contributions.

40 AI BASELINE

The fully automated work: sourcing candidates from Fellow social media, generating draft text, suggesting categories and tiers, and structuring the raw report. This contribution exists regardless of what the human editor does.

60 HUMAN EDITORIAL — MAX

The maximum human contribution, allocated across three dimensions of editorial judgment. The score reflects how much of this 60-point potential the editor actually exercises.

Component	Max points	Description
AI baseline	40	Sourcing, drafting, suggesting — always present
Selection Filtering	27	Reviewing and filtering AI-generated candidates
Tier Decisions	9	Prioritising entries as Featured or Top Story
Text Editing	24	Rewriting and refining AI-generated text
Total	**100**	

Traffic-Light Rating

The human score (0–60) maps to a rating that signals editorial health.

Active editorial oversight — `score ≥ 27`

Healthy human engagement across filtering, prioritisation, and editing.

Light-touch editing — `15 – 26`

Low engagement on one or more dimensions; worth reviewing editorial practice.

Review needed — `score < 15`

Very low engagement; output may not reflect meaningful human judgment.

Dimension Calculations

1. Selection Filtering — 27 points max

```
Input    sc-camel-skip-rate = skipped candidates / total candidates reviewed
Formula   min(skipRate / 0.35, 1.0) × 27
```

Logic. A skip rate of 0% means every AI-generated candidate was published without filtering — a rubber-stamp signal. The formula gives proportional credit up to a 35% skip rate (full marks), recognising that even modest filtering (20%+) shows active editorial judgment. Skip rates above 35% still score full marks; the additional filtering doesn't increase the human contribution signal.

Note on report quality. As the AI report improves through feedback loops, skip rates should naturally decrease (fewer irrelevant candidates). The 35% threshold is set low enough that a well-tuned AI pipeline with a 20% skip rate still scores 15.4 out of 27 — substantial credit for active review even when most candidates are publishable.

Current status. Skip rate is estimated at 48% based on manual analysis. Future versions will track this dynamically via the fact-checker tool.

2. Tier Decisions — 9 points max

```
Input   sc-camel-tiered-rate = entries tagged Featured or Top Story / total published
Formula   min(tieredRate / 0.5, 1.0) × 9
```

Logic. Each weekly digest has one Featured story and up to 3–4 Top Stories. If all entries remain at the default "Regular" tier, no prioritisation judgment is happening. The formula gives proportional credit up to 50% tiered entries.

Limitation. The AI currently suggests tier assignments. Since we do not yet track whether the human accepted or overrode the AI's suggestion, the tier score may overstate human contribution. This dimension carries the lowest weight (9 points) to reflect this uncertainty.

3. Text Editing — 24 points max

```
Input   avgEditPct across all entries with tracking data
Formula   avgEditPct / 100 × 24
```

Logic. `editPct` is a character-level edit distance between the AI-generated draft and the published version, recorded per entry:

- **0%** — AI text accepted verbatim
- **10–20%** — moderate editing (typical range)
- **50%+** — heavy rewriting or substantial restructuring

The raw average is used directly as a proportion of the 24-point weight. No normalisation cap is applied — higher editing effort contributes proportionally.

Important context. Even a 100% rewrite starts from the AI draft as a reference. The editor reads the AI output, decides it needs changing, and writes new text informed by (or reacting to) the AI's framing. This is why text editing can contribute at most 24 points, not the full 60 — the AI's draft is always the starting point.

Why 60 Is the Maximum

Every dimension of TxHAPulse production involves AI:

- **Selection filtering** chooses from an AI-generated candidate pool. The human did not source the stories.
- **Tier decisions** confirm or override AI-suggested tiers. The AI provides a starting point.
- **Text editing** revises AI-generated drafts. Even full rewrites are informed by the AI's framing.

A score of 100 would require the human to source stories independently, write from scratch, and make every structural decision without AI input. That is not how TxHAPulse works, nor is it the goal. The 40-point AI baseline acknowledges that AI does substantial work — and that this is by design.

In practice, human scores in the **27–45 range** indicate healthy collaboration — the human is making substantive editorial decisions at every stage while leveraging AI for the work it does well.

Where the Score Appears

DIAGNOSTIC DASHBOARD

`diagnostic.html`

Full breakdown with per-dimension scores, progress bars, raw inputs, and a methodology link. Headline displays as **"X / 60."**

FACT-CHECKER

`fact-checker.html`

A warning banner appears during review when the live score drops to yellow or red, prompting the editor to increase engagement before committing. Hidden when the score is green.

Future Improvements

1. **Dynamic skip tracking.** The fact-checker will persist skip counts per week, replacing the current estimated skip rate with actual data.
2. **Tier override tracking.** Recording whether the editor changed the AI's suggested tier, giving the tier dimension a more accurate signal.
3. **Per-week trends.** Showing how the score evolves over time to surface gradual drift toward rubber-stamping.
