Neutral Is Not Free

Evaluating Downside Risk in Neutral Launches

Pablo Alcain · Jason Kang · Mia Garrard · Marine Veits · Houssam Nassif · Abbas Zaidi
Meta

arXiv:2610.10223

1 “Neutral” hides very different risks

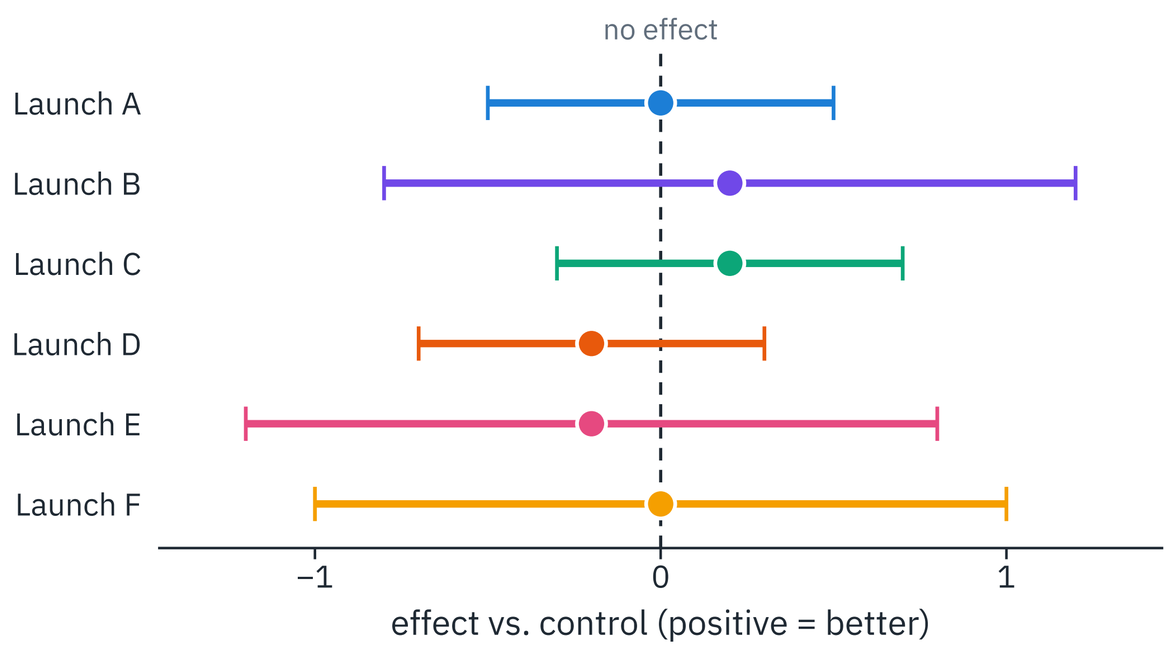
Migrations, dependency upgrades, reliability and cost work ship for benefits the metric does not see. Review asks only whether they hurt it, usually by checking that the confidence interval contains zero.

The interval rule fails twice

It rewards noise. The interval contains zero whenever $|\hat{\theta}| \leq 1.645\,\mathrm{SE}$, so noisier estimates pass more easily: the rule is permissive with scarce data and strict with abundant data.

It is binary. “Neutral” says nothing about how much the metric could lose.

> **Absence of evidence is not evidence of absence**



Six synthetic results. All six 90% intervals contain zero, so the usual rule calls every one of them neutral.

2 Report both PoF and EBL

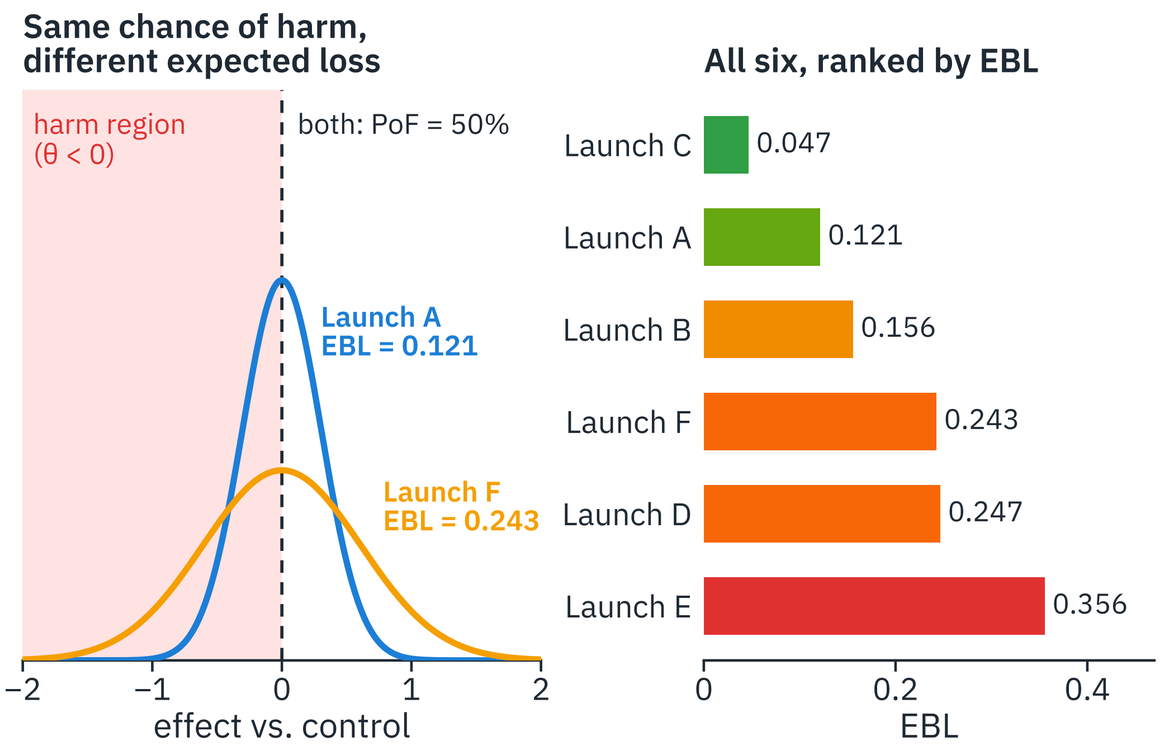
$$\mathrm{PoF} = P(\theta < 0 \mid D)$$

How likely it is to be negative

$$\mathrm{EBL} = \mathrm{PoF} \times \big(-\mathbb{E}[\theta \mid \theta < 0,\, D]\big)$$

How negative the negative is, in the metric’s units

θ: true effect on a guardrail metric (larger is better); D: the experiment’s data. EBL is the posterior expected loss of launching, $\mathbb{E}[(-\theta)_+ \mid D]$.



Same PoF, different EBL. A and F both have PoF = 50%, but F is twice as uncertain, so its EBL is twice A’s.

From any A/B readout

Flat prior, normal posterior with $m = \hat{\theta}$ and $s = \mathrm{SE}$:

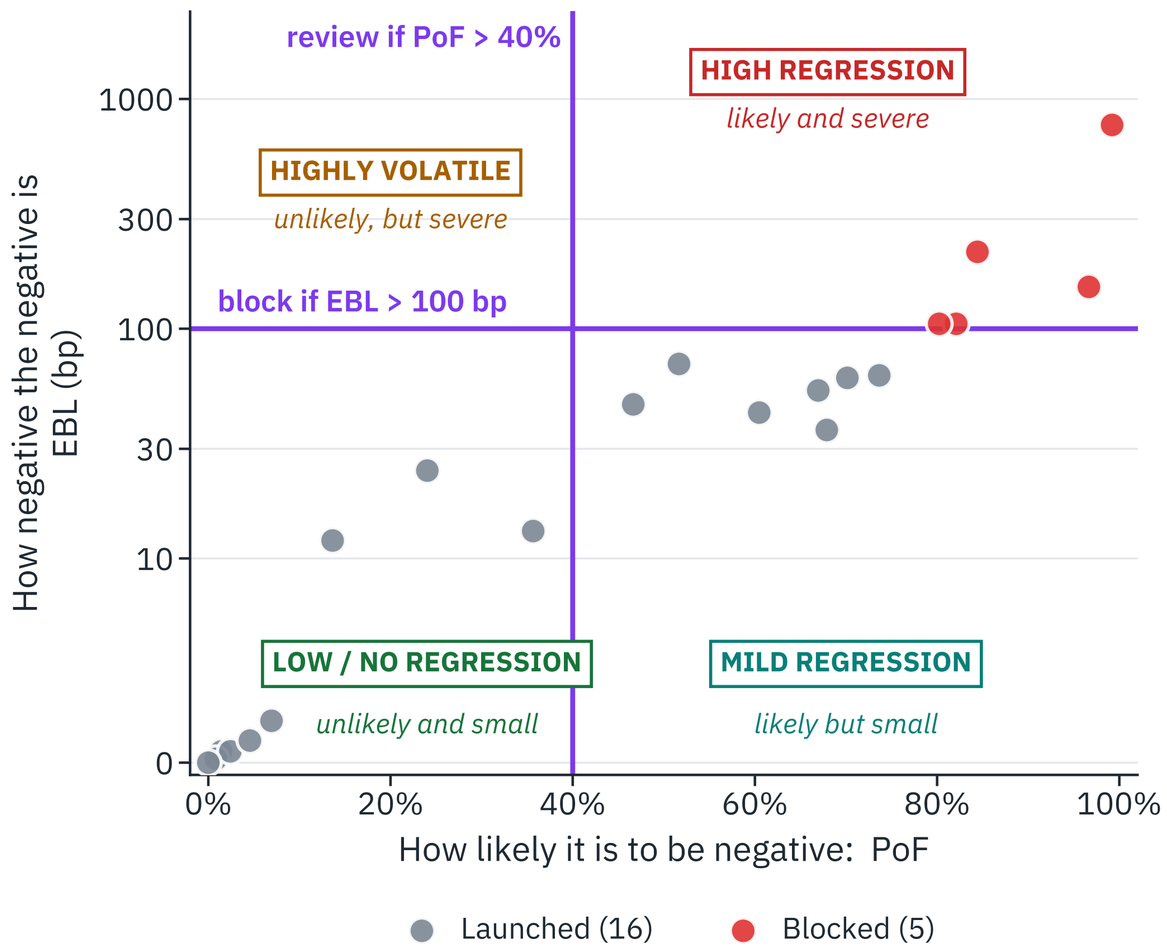
$$\mathrm{PoF} = \Phi\!\left(-\tfrac{m}{s}\right), \qquad \mathrm{EBL} = s\,\phi\!\left(\tfrac{m}{s}\right) - m\,\Phi\!\left(-\tfrac{m}{s}\right)$$

From a 90% interval $[L, U]$: $m = (L+U)/2, \quad s = (U-L)/3.29$

Why both. With $u = m/s$, $\mathrm{PoF} = \Phi(-u)$ ignores the noise s, while $\mathrm{EBL} = s\,[\phi(u) - u\,\Phi(-u)]$ grows with it. PoF says how likely a regression is, EBL how much is at stake; together they separate a likely-but-small regression from an unlikely-but-severe one.

3 PoF and EBL match expert decisions

21 expert-reviewed experiments on one guardrail metric: 16 launched, 5 blocked. EBL in bp (100 bp = 1% lift).



Blocked launches sit above 100 bp, approved ones below. PoF splits likely-but-small (bottom right) from unlikely-but-severe (top left).

16/16
launches below 100 bp (also leave-one-out)

5/5
blocks above 100 bp (also leave-one-out)

>95%
review agreement (paper) (CI lower bound: <80%)

PoF at 80% separates these too: s varies little here, so the scores agree; they diverge when precision differs.

Using PoF and EBL in review

block	**EBL above 100 bp**
review	**PoF above 40% or EBL above 30 bp**
launch	**otherwise**

The block threshold states a risk appetite: the expected loss this class of launch is worth. Score $(-\theta - \Delta)_+$ to tolerate losses up to Δ.

21 experiments: thresholds are candidates to validate on new launches. EBL weighs the left tail, so it is sensitive to a misspecified posterior.

Altman & Bland (1995) *BMJ* · Berger (1985) *Statistical Decision Theory and Bayesian Analysis*