

Feedback Loops

Communications & Media

A system trained on its own output stops measuring what it set out to.

How the loop closes.

1

Rank by predicted response

Items with a higher predicted click or watch are shown more.

2

Exposure creates data

Only shown items get feedback, so unshown items stay unproven.

3

Retrain on that data

The model learns from a sample its own choices produced.

4

Behaviour adapts

Creators optimise for the signal, so the signal changes meaning.

5

The metric decouples

The number rises while the outcome it proxied does not.

Goodhart's law in a machine: once a measure becomes a target, it stops being a good measure. Holdouts and off-policy tests are the usual