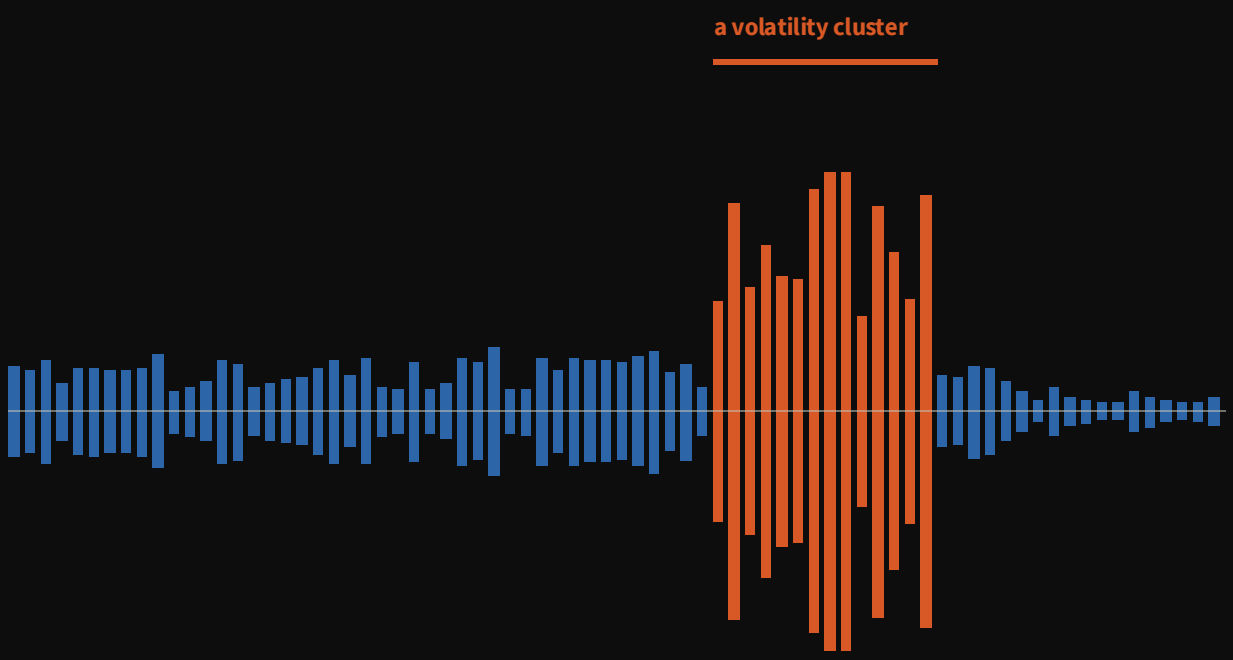
HANSE HEDGE
free information sharing for quants and algorithmic traders

WHITEPAPER



touch the floor once and the account is gone

Survival first.

Benchmarking trading algorithms under a **prop-trading survival mandate**

A proving ground for position sizing using Monte Carlo resampling, a CPPI risk cushion and a margin safety switch. The score asks who is still trading at the end. Factor models generate alpha, and position sizing operationalizes it.

Marius Ciepluch
marius.ciepluch@gmail.com
hanse hedge · August 2026

CONTENTS

What's here and where to find it.

> *AI will be used to generate alpha at scale. That's one side of the coin. Executing the strategy is the other, and many underestimate the operational challenges.*
>
> - Marius Ciepluch, M. Eng MBA CFE

START HERE

Four questions. If you read only four pages, make them these.

1 **Why should I care right now?** page 4

AI investment is large enough to reprice markets. The next few years may be more volatile, not calmer.

2 **What does "survival" actually mean?** pages 5-8

A pass/fail bar from prop trading: a floor you cannot touch, a daily loss limit and a profit target. Touch the floor once and the account is gone.

3 **How do you prove an algorithm survives it?** pages 9-17

Test it on resampled years that preserve volatility clusters, include real costs and require paper trading before any money moves.

4 **Where is this wrong?** page 23

The survival score is in-sample. Near the floor, the live rule takes more risk than the model assumes. Both limits are stated plainly.

WHAT THIS IS

A test bed and position-sizing rule. The pass/fail bar is fixed before testing, and known failures are listed at the back.

WHAT THIS IS NOT

It is not a signal or return forecast. It does not claim to predict the next volatile stretch.

Your role: you do not run the prop account or monitor the optimiser. You receive an algorithm that has already passed through the test bed, along with its record.

WHY NOW

AI investment is the largest capital reallocation in a generation. Plan for more market noise, not less.

01

Capital concentrates

Huge sums flow into one theme, often with borrowed money and long payback periods.

02

Positioning crowds

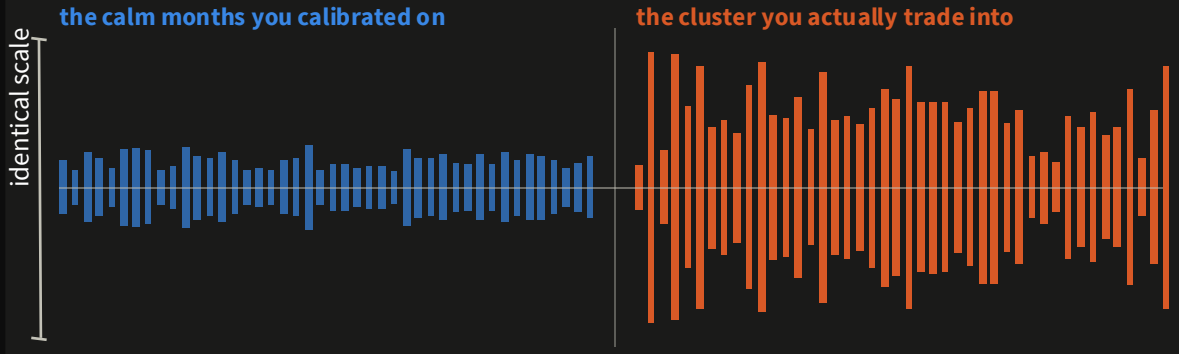
Investors end up in the same few names and hedges, sized from calm historical data.

03

Shocks spread faster

When the theme wobbles, everyone heads for the same exit. Calm periods can end abruptly and cluster.

WHAT A SIZING RULE IS CALIBRATED ON, AND WHAT IT MEETS



Schematic: daily moves drawn on one shared vertical scale, no units implied. Real figures start on page 18.

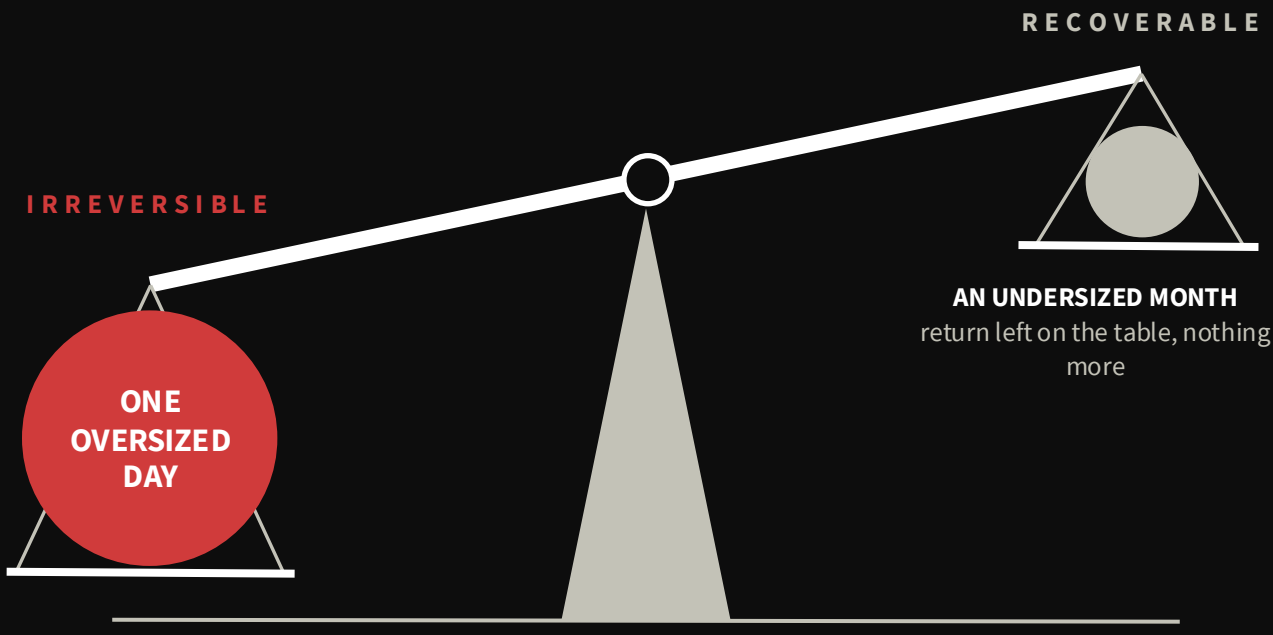
The rule and size are unchanged, but the market is not. There is no warning at the boundary. That is the problem.

The useful question is not "Which algorithm returned the most last year?" It is "Which algorithm is still trading after a bad one?"

THIS IS WRONG IF

the next few years are calmer than the last. Then this method gives up some return for protection that was not needed. That is the trade: a small cost against the asymmetric risk shown on the next page.

One oversized day ends the account. An undersized month only costs return.



A levered book with a survival mandate has an asymmetric loss function. Sizing must price that asymmetry rather than average it away.

The account can recover from taking too little risk. It cannot recover from one day of too much. Growth-optimal sizing and volatility targeting account for this only partly because neither includes a hard floor.

The rest of this handout follows from treating that asymmetry literally.

If ruin is the risk, return alone ranking is the wrong test. Risk management matters

A ranking sorts algorithms by what they made. It does not care about the single day that ends the account because that day disappears into an average.

The test needs a bar: a fixed line that the algorithm either stays above or crosses. Prop trading firms already use one, so this framework borrows it.

RANKING MINDSET

"It beat the benchmark by 4 points."

Rewards the size that paid off in that sample. Ruin appears as one bad month in the average.

BAR MINDSET

"It never touched the floor in any tested year."

Pass or fail comes first. Returns are compared only among the algorithms that passed.

Next: the bar itself. Three numbers and one daily limit decide whether the algorithm is still in the game.

Example mandate. It is simple geometry: equity must stay between EUR 144,000 and EUR 184,000.

ACCOUNT EQUITY, EUR (basic example)

The floor is the problem statement, not a preference. The sizing law follows from this geometry.

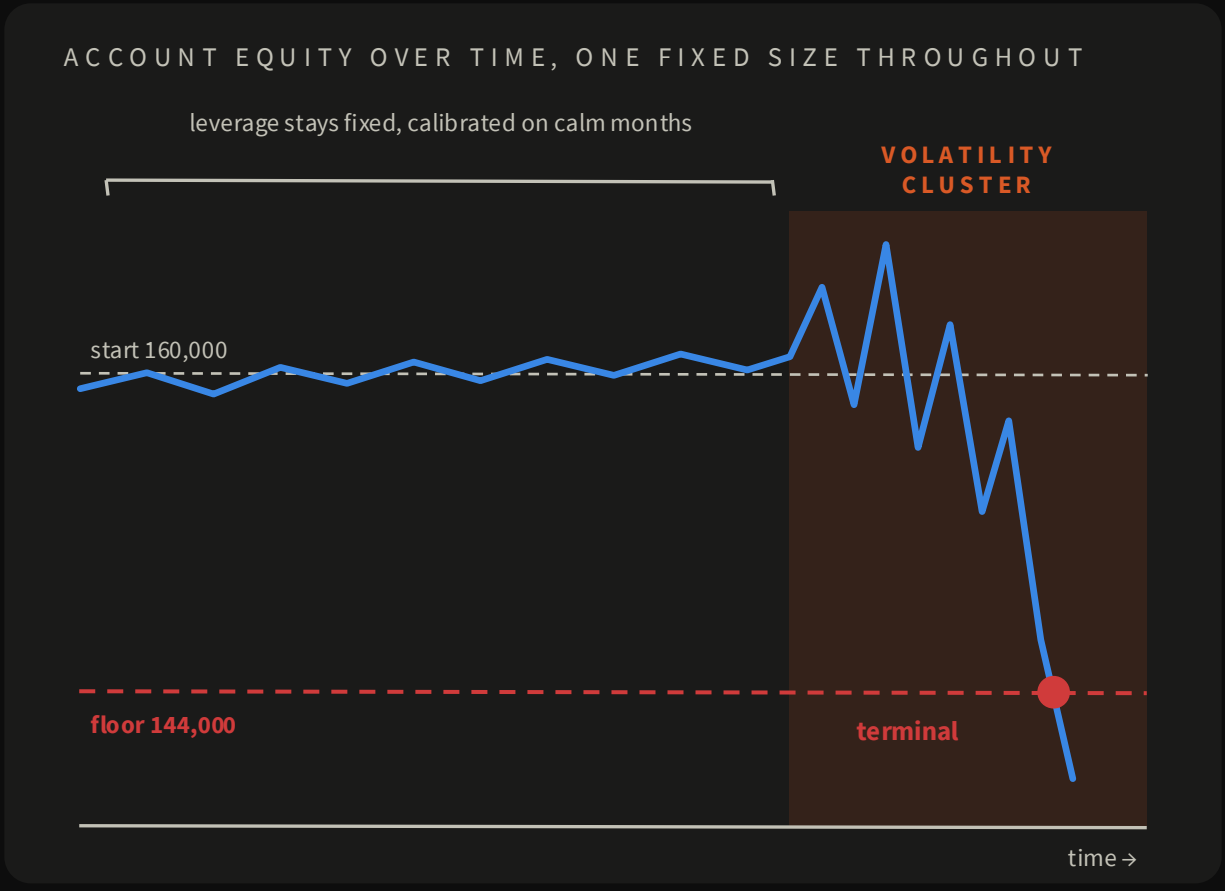
ANCHOR 1 · THE CUSHION

Risk comes from the distance to the floor. The cushion, equity minus 144,000, is the only source of risk budget (page 12).

ANCHOR 2 · THE DAILY RUNG

The EUR 8,000 daily limit is the denominator for every intraday stress gauge later in the whitepaper (page 16).

What usually goes wrong: fixed leverage meets a volatility cluster.



Volatility clusters. The account is most likely to fail just after recent history looked calm.

- **The size is largest when it matters most. A position calibrated for a calm regime is too large for the cluster, which arrives without warning.**
- **The reaction is too late. One bad day in the cluster can breach the daily limit before de-levering begins.**
- **The fix is structural rather than predictive. The sizing rule must assume clusters and reduce leverage before they arrive. It does not need a forecast (page 10).**

An algorithm earns capital only after surviving the mandate at all three stages.

THREE STAGES, EACH WITH A GATE

RESAMPLED YEARS
Gate: did it touch the floor in any of them?

Hundreds of alternate versions of the recent past, rebuilt so calm and violent periods stay together.

PAPER TRADING
Gate: does live plumbing change the result?

The same code, data feed and order interfaces, but no real money. Latency, rounding and financing still apply.

LIVE, UNDER THE MANDATE
Gate: still trading at the end?

A funded evaluation account with a hard floor, daily limit and target. The account is the referee.

WHAT THE TEST BED REFUSES TO LEAVE OUT

- **Spread, commission and financing** — charged on every rebalance, including overnight swap
- **Lot rounding and minimum size** — a budget below minimum size means no trade, not a small trade
- **Margin and leverage limits** — the broker may refuse before the model stops
- **Clustered volatility** — resample in blocks so bad days stay together
- **The daily loss limit** — an intraday rule, not an end-of-day average
- **One data source, named** — FMP prices; the method does not depend on the vendor

The method is vendor neutral. It models a class of FTMO-style compound evaluation accounts, not a specific product, broker or platform. The frictions above are explicit rather than assumed away. If an algorithm fails a gate, it goes back. No stage can be skipped.

The framework clears the bar. The dial selection does not, and the results say so.

01

The sizing framework clears the mandate

50-61%

of one-year mandate windows pass across the full dial range

Cushion law, safety switch and throttles combined. For context, the commonly cited prop-firm evaluation pass rate is about 14%.

02

The improvements that held up are simple controls

0.429 → 0.229

holdout fail rate for the control set chosen on development data

A plain cap of the daily risk budget at the daily loss limit does most of the work on its own (holdout pass 0.714).

03

The dial-selection layer shows no detectable benefit

0.556 vs 0.611

adaptive versus best fixed dial; intervals overlap

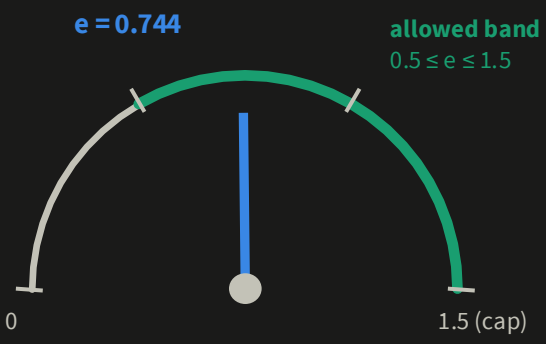
It also keeps the incumbent on 67.6% of recalibrations. That is an honest result, not optimiser alpha.

Scenario: AS-DEPLOYED. This is the rule as the live binding runs it. Momentum and correlation throttles are off because live hard-codes both to 1.0. Page 19 tests them on.

What you receive is not a clever dial. It is a rule that survives honest measurement, a short list of controls that passed the holdout and a record of what failed.

Risk appetite is one dial built from two inputs: exposure e = m × r / 100, capped at 1.5.

ONE NUMBER, TWO KNOBS



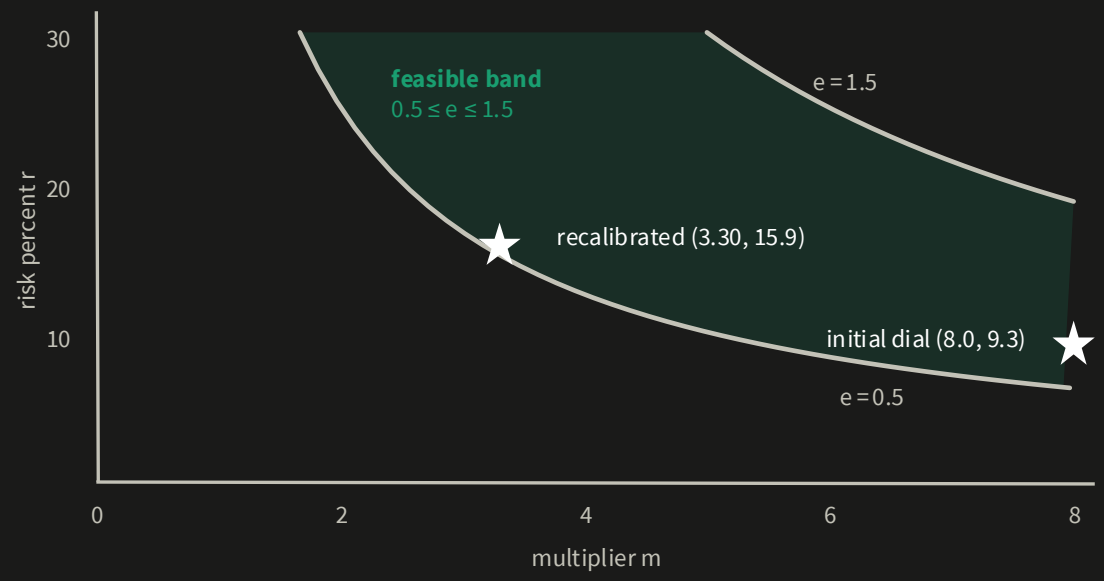
e = m × r / 100

m, multiplier: gain applied to changes in the risk cushion

r, risk percent: share of the cushion put at risk

A point in the (m, r) plane fully defines risk appetite.

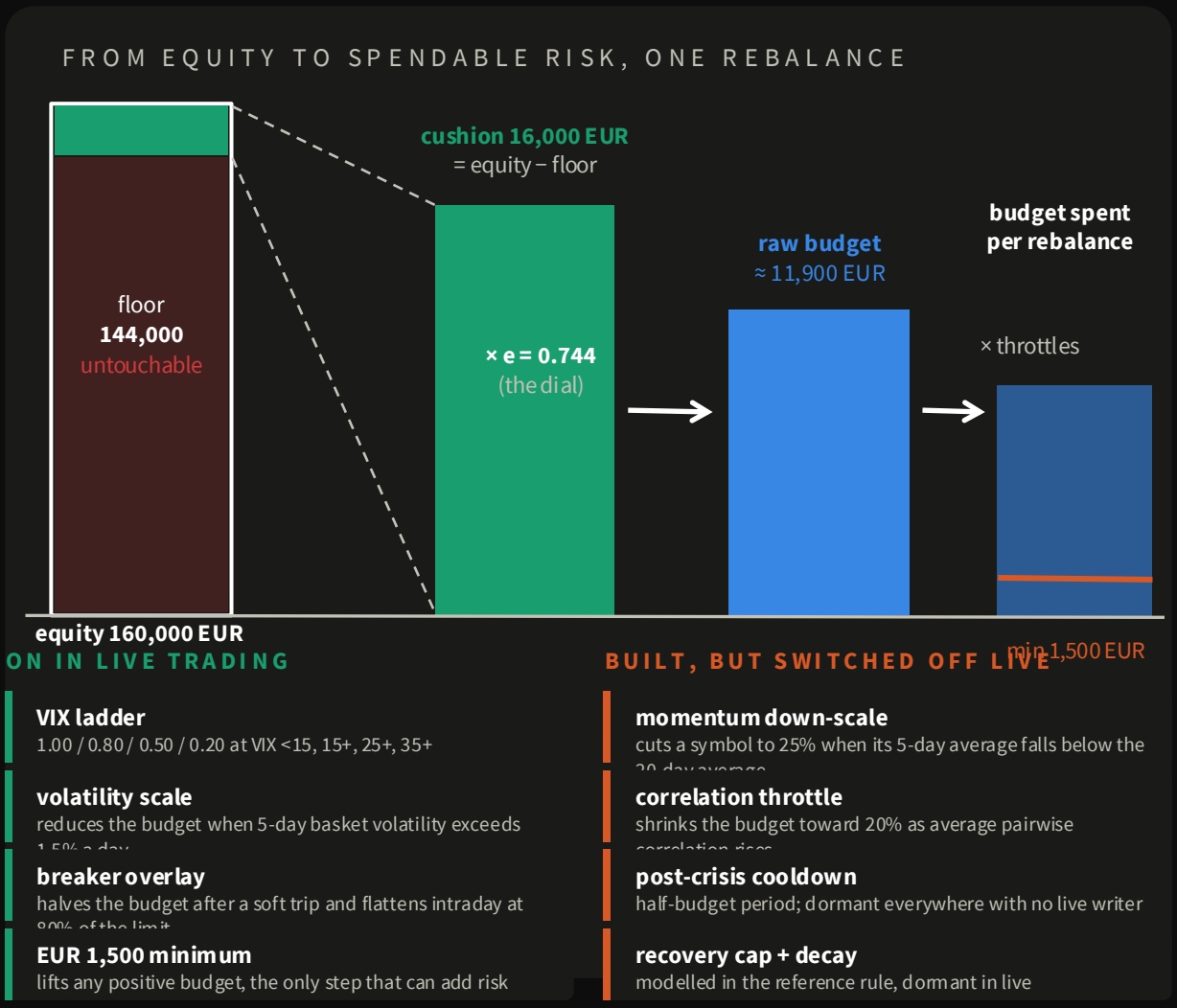
THE DIAL PLANE: EVERY POINT IS A CANDIDATE APPETITE



Equal exposure does not mean equal behaviour: a low m reacts more gently when the cushion jumps.

The 1.5 cap and 0.5 floor bound the search before optimisation. Page 15 shows why low and high values of m behave differently. The simulation finds that difference on its own.

Risk is spent from a cushion: budget = cushion × dial × throttles.



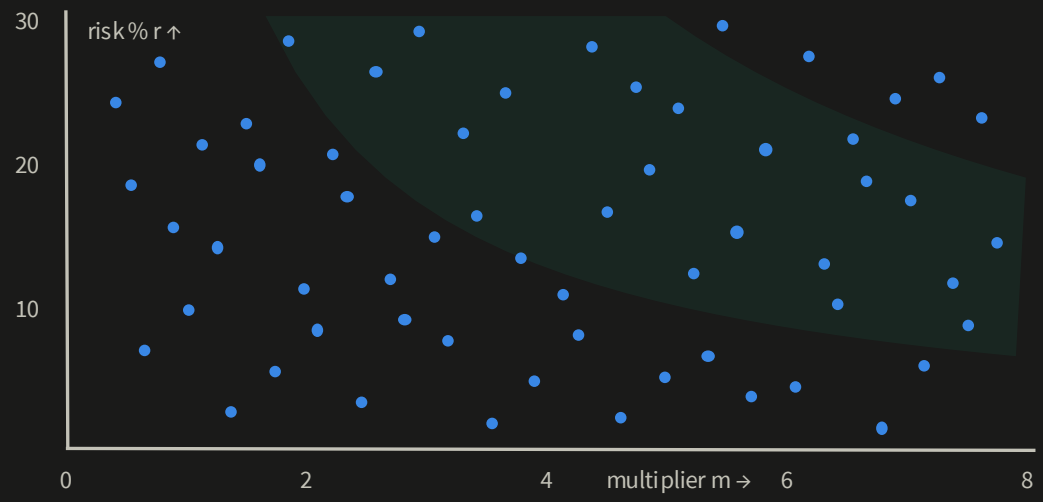
The cushion cannot exceed its initial EUR 16,000, so gains never increase risk. It falls to zero when the day's loss approaches the EUR 8,000 daily limit.

This is CPPI. Risk comes only from headroom above the floor. With no cushion, there is no risk budget. Losses cut size linearly; gains do not increase it.

The dial from page 11 is the only risk-appetite parameter in this chain, except for the EUR 1,500 minimum that prevents small budgets from rounding every position to zero.

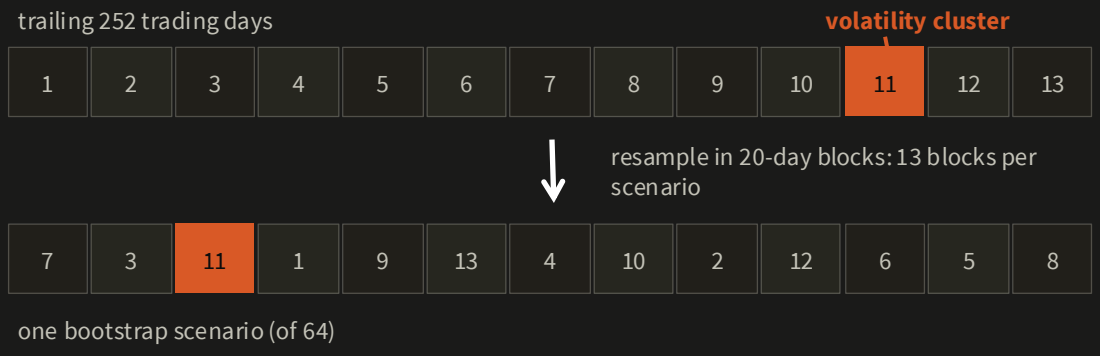
Every candidate dial is rehearsed on the same 64 resampled years, with the clusters left intact.

128 CANDIDATE DIALS, EVENLY SPREAD



The candidate set stays fixed at each monthly recalibration. Only the data window moves.

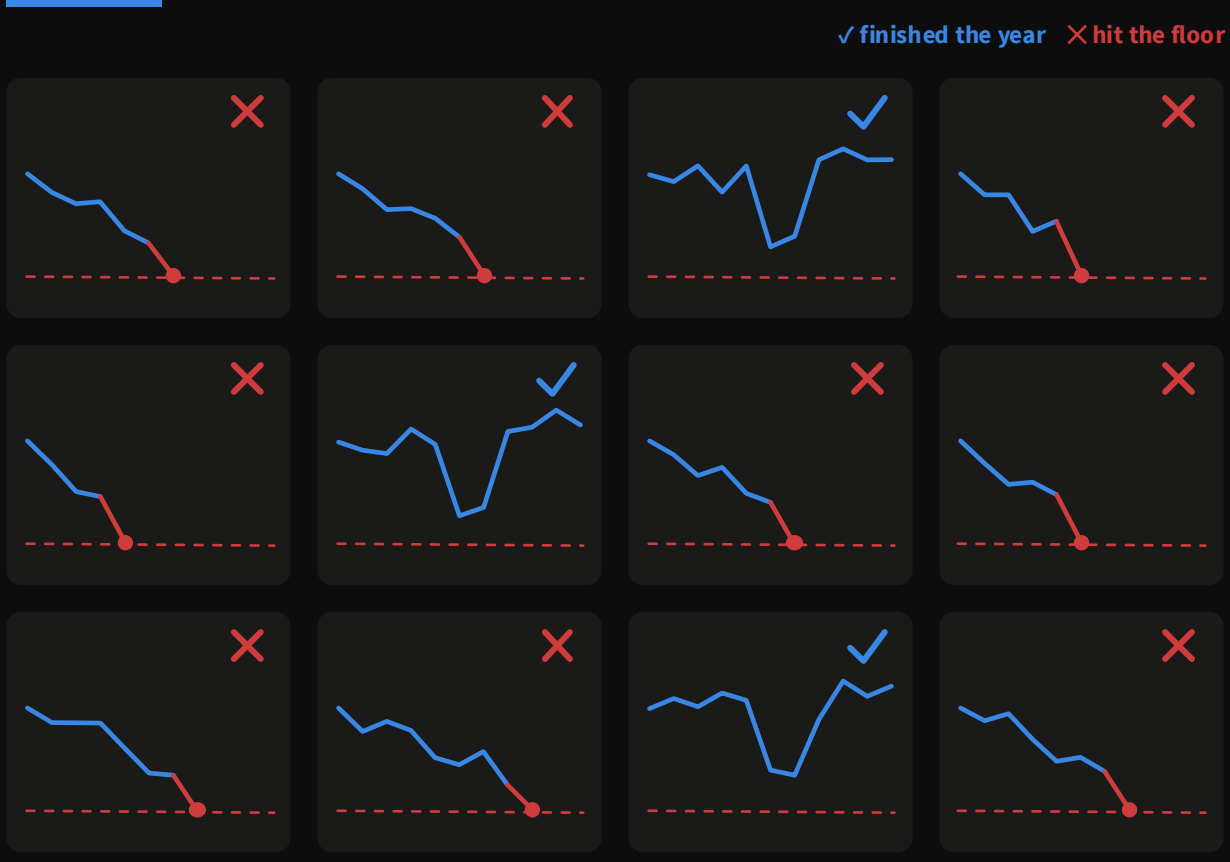
THE TRAILING YEAR, RESHUFFLED IN 20-DAY BLOCKS



Clusters, autocorrelation and crisis correlation stay inside the blocks. Reshuffling changes the order of regimes but does not remove them.

Each of the 128 candidate dials runs through the same 64 resampled years. Common random numbers give every dial the same draws, so luck does not affect the comparison.
Day-by-day resampling would break the clusters and create unrealistically mild years. Fixed seeds make each monthly recalibration exactly reproducible.

Most candidate settings fail. That is the result, not a flaw in the test.



Each rehearsed year uses the rule that ships: cushion, exposure cap, throttles, internal safety switch, spread, commission and financing. A floor breach or daily-limit breach ends the run, regardless of earlier returns.

0-30%

typical share of allowed settings that survive a resampled year; the survival tiers are never reached

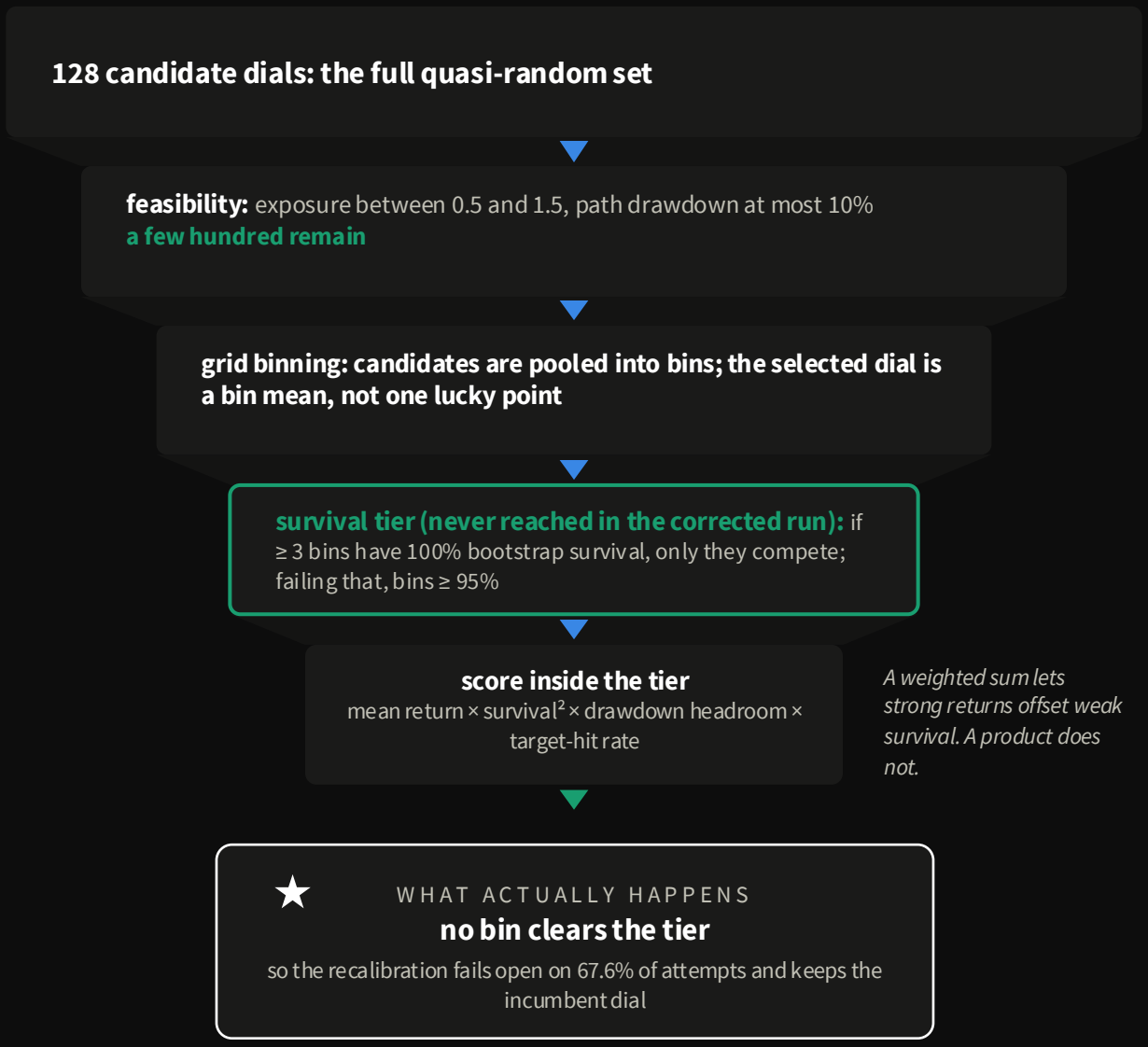
67.6%

of recalibrations find nothing that clears the gate and keep the previous setting

Most recalibrations produce an honest answer: "no improvement available." The system keeps the incumbent rather than choosing the least bad option.

SCENARIO: AS-DEPLOYED, THE ISOLATED RUN · 128 dials × 64 common scenarios, common random numbers; momentum and correlation throttles off, matching live.

The ranking layer is sound on paper but rarely runs in practice.

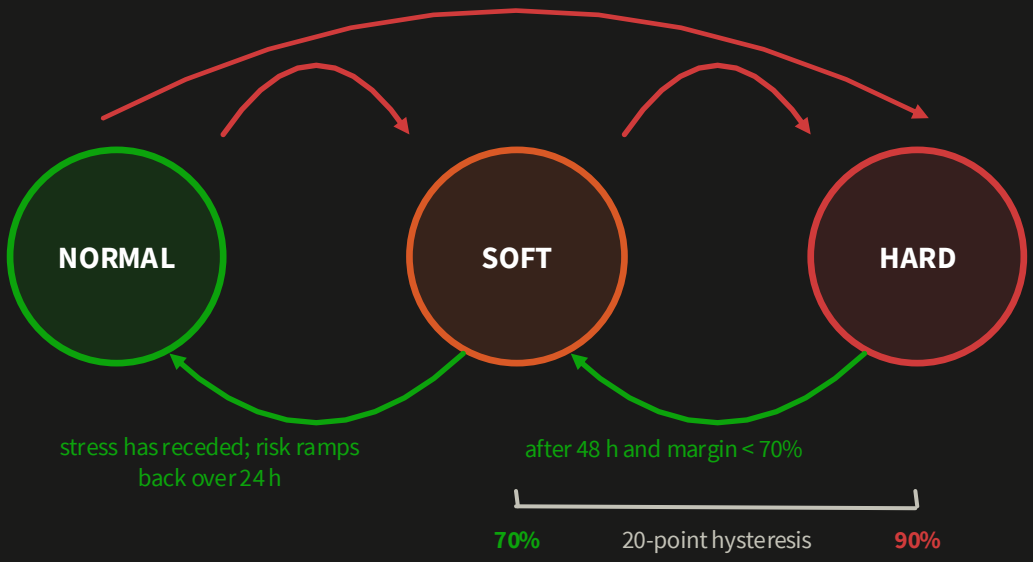


The design is strict by construction. Survival gates entry to the score and is squared inside it, so the penalty for ruin is quadratic rather than averaged away.

That strictness has a cost. Under the corrected estimator, bootstrap survival in the allowed range is 0-30%. The tier thresholds are never met, so the score is rarely computed. Failing open is safer, but it leaves the ranking machinery inactive most months. That is why the dial layer shows no measurable edge on page 20.

The margin safety switch bounds whatever the optimiser gets wrong: trip at 90%, re-enter below 70%.

THREE STATES, ONE-WAY ESCALATION AND SLOW RE-ENTRY



NORMAL full risk budget

SOFT budget halved; trips when the day's loss reaches 65% of the EUR 8,000 limit

HARD book flattened at 90% margin utilisation or 80% of the daily limit

Separate hard cap: gross notional cannot exceed 1.34 × equity. ATR-based sizing does not see notional, so this is the binding systemic control.

Margin enters the system at one decision point: this switch. The sizing formula never reads margin, which keeps it a pure function of equity and the dial.

Hysteresis is intentional: trip at 90, re-enter below 70 and wait 48 hours. This prevents oscillation at the boundary.

The simulation chooses the appetite, the control law spends it and the safety switch bounds it. None can override the others.

"You picked the dial on the recent past. So what?"

Fair point. Everything so far was selected by resampling a trailing year, a procedure that can flatter itself.

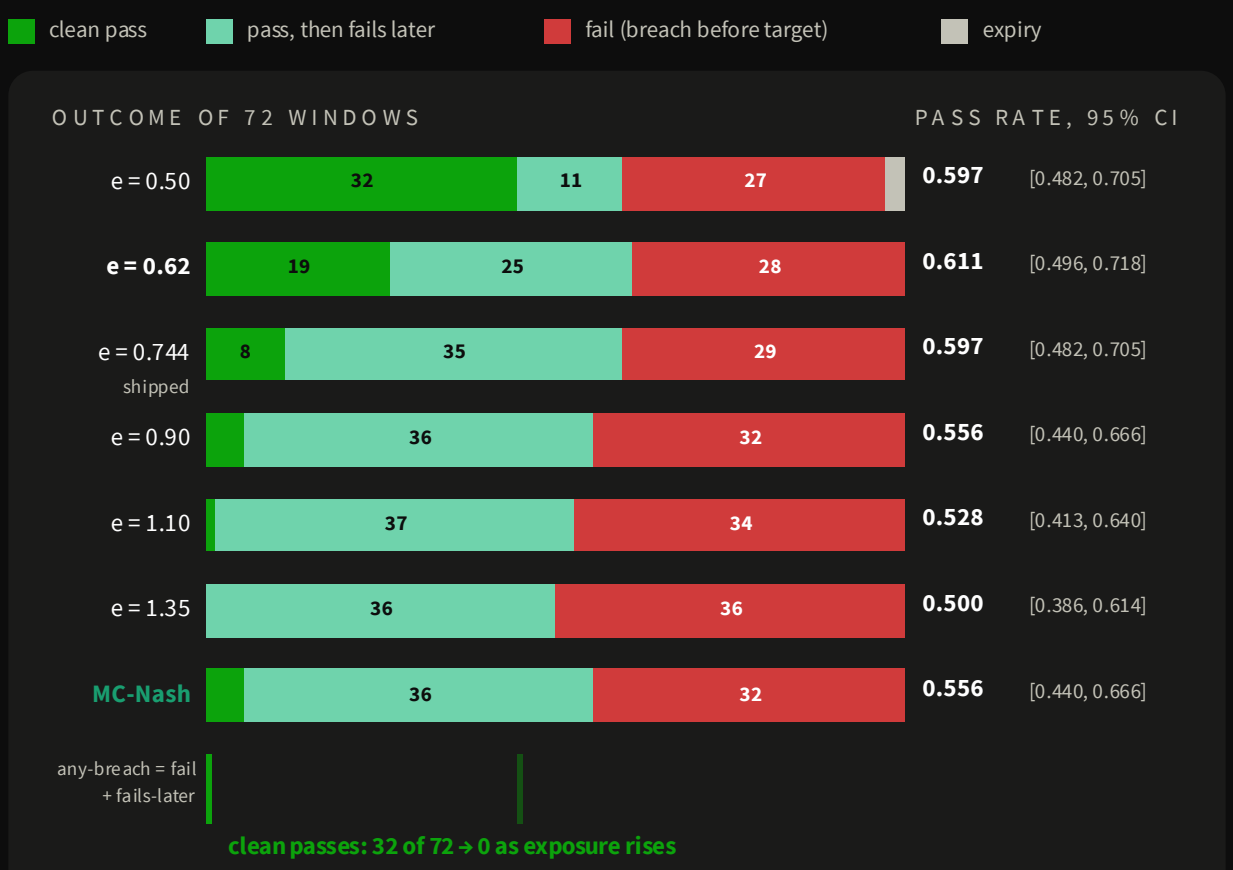
The rule is therefore replayed across two decades of overlapping one-year windows. It reselects the dial each month using only information available at that time. Every contender uses the same control law, costs and mandate.

WHAT TO LOOK FOR ON THE NEXT PAGE

- **The outcome is not return. It is whether the account finished the window still trading.**
- **Do not expect a large change in pass rate. Across the usable exposure range, pass rates are close and the confidence intervals overlap.**
- **Exposure changes the quality of a pass. It determines whether an account stays passed or reaches the target and fails later.**

The claim is modest. The method selects a survivable exposure level without hindsight and keeps the account there. It does not forecast volatility.

Pass rates are flat across the dial. What exposure changes is how you pass.



Flat pass rate

All pairwise intervals overlap. Lower exposure does not raise the pass rate.

Pass quality changes

Higher exposure turns clean passes into pass-then-fail windows: the account reaches the target and later gives it back.

The adaptive dial ties

0.556 versus 0.611 for the best fixed dial. Across paired windows, the adaptive dial is better in 3 and worse in 7 (descriptive).

Scenario: AS-DEPLOYED, the isolated run · 128 dials × 64 common scenarios, common random numbers, opening-gap execution checks, four-way absorbing outcomes. Jeffreys 95% intervals; windows overlap, so any p-value is descriptive only.

Read the bars rather than the averages. When the green segment shrinks and the pale segment grows, the account still reaches its target but later gives it back.

Switching on the full throttle stack makes every dial safer. It requires one configuration change.

Two runs use the same build and estimator. The isolated run matches live behaviour; the implementation run turns on the full designed stack.

n = 72 windows	AS-DEPLOYED isolated run	AS-DESIGNED implementation run	delta
fixed e = 0.62, pass	0.611	0.653	+0.042
fixed e = 0.62, fail	0.389	**0.278**	**−0.111**
fixed e = 0.62, clean passes	19 (26%)	**35 (49%)**	**× 1.8**
shipped e = 0.744, fail	0.403	0.375	−0.028
shipped e = 0.744, clean passes	8 (11%)	**16 (22%)**	**× 2**
MC-Nash, pass	0.556	0.556	0
MC-Nash, clean passes	4 (6%)	**14 (19%)**	**× 3.5**
MC-Nash, declines to act	67.6%	57.8%	−9.8 pts

Both sides use the same estimator: 128 dials × 64 common scenarios, opening-gap execution and four-way outcomes. The isolated run fixes momentum and correlation at 1.0, matching live. All intervals overlap.

THE STACK REDUCES FAILURES

At the tuned dial, the fail rate falls from 0.389 to 0.278 and clean passes nearly double. Every dial moves in the same direction.

THE OPTIMISER DOES NOT IMPROVE

MC-Nash passes 0.556 in both configurations. The stack makes every dial safer, but it does not choose a better one.

The simplest finding is also the cheapest to implement. Momentum and correlation throttles already exist in the code but are hard-coded to 1.0 in live. Turning them on is a configuration change. In this sample, it doubles clean passes at the shipped dial.

The point estimates are sizable and consistent in direction, but not proven. The intervals still overlap across 72 overlapping windows.

Measured on its own, the Monte Carlo dial layer adds nothing detectable.

This page asks one question. With no alpha, momentum throttle or correlation throttle, which is the live configuration, does simulation-based dial selection beat a fixed dial?

HEAD TO HEAD, 72 WINDOWS

fixed dial, e = 0.62

0.611

pass rate [0.496, 0.718]

adaptive MC-Nash

0.556

pass rate [0.440, 0.666]

Across the same windows, the adaptive dial does better in 3 and the fixed dial in 7 (p = 0.344, descriptive only). Against the shipped default, the count is 2 versus 5.

67.6%

of recalibrations find nothing that clears the gate and keep the incumbent dial

20 of 72

windows in which the dial never moves from the incumbent at all

9 of 72

windows where it beats the shipped default on absorbing return

TWO BORROWED IDEAS THAT TURNED OUT INERT

- **Survival tiering (LCB): identical to baseline. Bootstrap survival never reaches the thresholds, so the tier never engages.**
- **Volatility-conditioned sampling: no effect on the holdout. The earlier claim that it improved all outcomes is withdrawn.**

For scale, median absorbing return is about +15.2% to +15.4% for every dial at or below 0.90. The safety switch fires about 2.4 to 3.3 times per simulated year at every dial.

> **Why publish a negative result? Because measurement is the product. An earlier version reported a large optimiser edge, but the corrected estimator removed it. Publishing that correction makes the remaining numbers worth reading.**

The account blew its daily limit at the opening bell. Zero safety-switch trips.

After a strong US payrolls report, GLD and TLT gap down together at the open. One move hits both legs.

daily loss limit

EUR 8,000

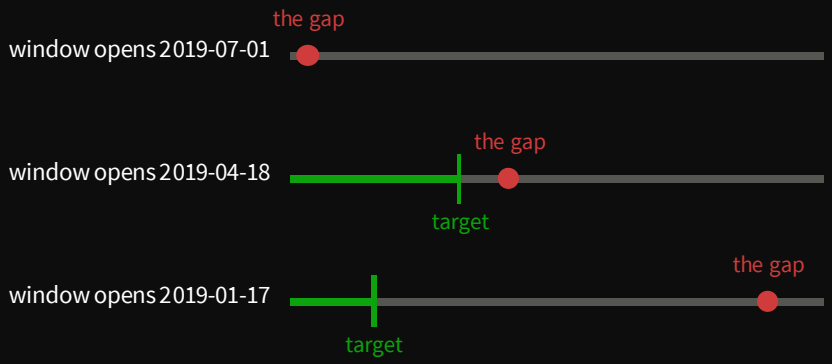
realised on the gap, at the shipped dial e = 0.744

EUR 8,096

EUR 8,094 comes from the price gap. No trading decision occurs between the close and the open.

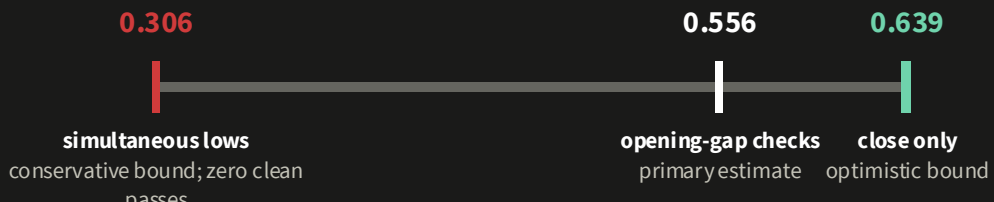
0 switch trips

THE SAME GAP, THREE EVALUATION WINDOWS, TWO DIFFERENT VERDICTS



The event is identical. Whether it ends the account or becomes a footnote depends only on the window's starting point.

HOW EXECUTION ASSUMPTIONS CHANGE THE ADAPTIVE DIAL PASS RATE



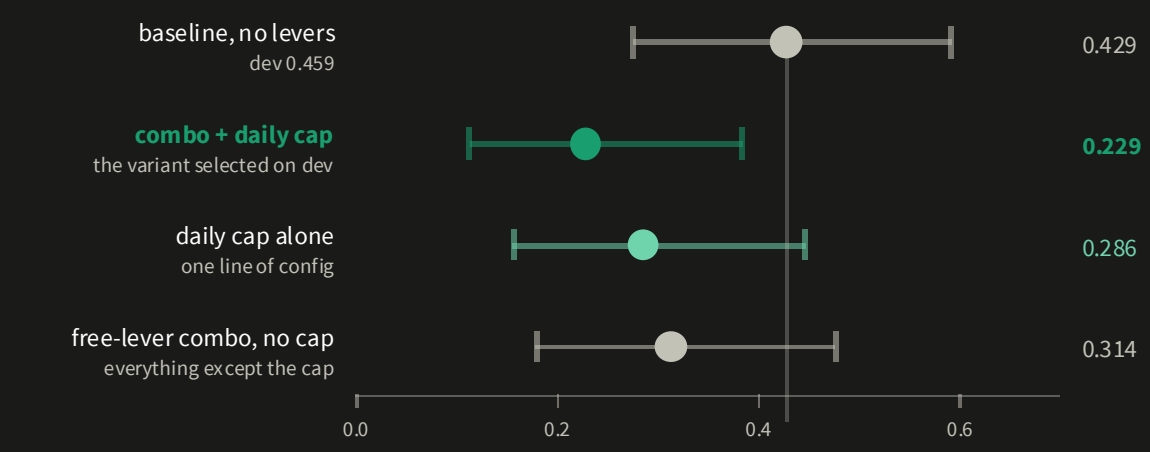
Daily data cannot narrow the range further. The true rate lies somewhere inside it.

The missing control is overnight exposure. An intraday circuit breaker cannot stop a move that occurs while the market is closed. Tuning the dial will not fix this. The strategy needs a rule for what the book may hold overnight.

The only holdout improvement is also the simplest.

The levers were selected on 37 pre-2017 windows, then tested unchanged on 35 windows from 2017 onward. Fifteen variants were tried, and that multiplicity is part of the result.

FAIL RATE ON THE HOLDOUT: BREACH BEFORE TARGET, ADAPTIVE DIAL



Dots show point estimates; bars show Jeffreys 95% intervals. Every interval overlaps the baseline. The direction is consistent, but the difference is not proven.

0.429 → 0.229

holdout fail rate, baseline against the combination chosen on development data only

Cap the daily budget

Capping the risk budget at the daily loss limit does most of the work alone. Holdout pass rate: 0.714.

AND WHAT MADE IT WORSE

post-target de-risk alone	0.486
post-flatten hold alone	0.457

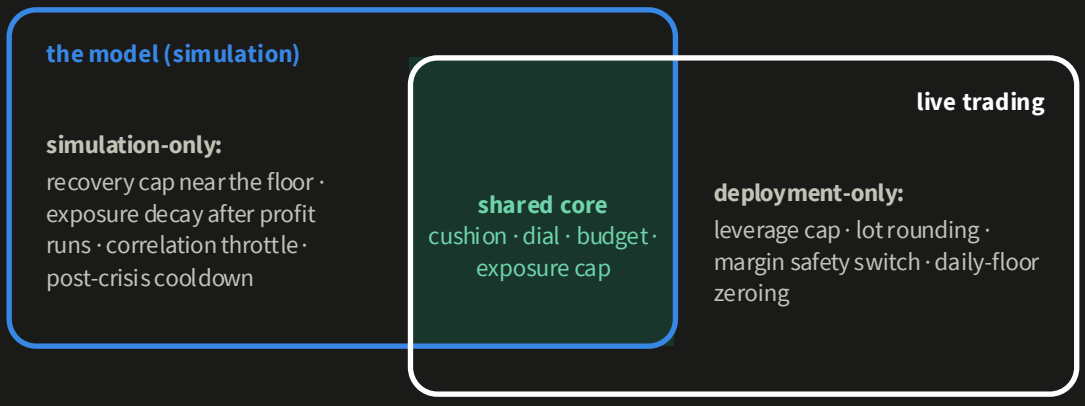
Both are worse than the 0.429 holdout baseline. They remain in the record.

The overall pattern is simple: basic constraints help, the clever layer does not, and only a holdout separates the two.

THE HONEST LIMITS

The model is not live trading, and every estimate has a range.

THE MODEL VS LIVE TRADING



The dial is optimised on an approximation of live trading. Near the floor, the shipped rule takes more risk than the model.

WHAT THE SCORE SAYS, AND WHAT IT DOES NOT

- The 72 one-year windows overlap, so every p-value is descriptive, never "significant".
- It says that no sampled year breached that dial region under the modelled rule.
- **execution assumptions alone move the pass rate from 0.306 to 0.639**

Known structural edge: within about 2,000 EUR of the floor, the 1,500 EUR minimum budget can risk more than the remaining cushion.

These caveats are part of the paper and part of the method. Adopting the method means accepting them.

The score is an in-sample bootstrap statistic. The real test is out of sample, in a volatility regime absent from the trailing year (page 18). The safety switch exists because the model will be wrong somewhere; it limits the cost of that error.

END OF THE WHITEPAPER

Take the method, test it yourself, and tell me where it breaks.

WHERE TO GO NEXT

The working paper
full derivation, estimator details and complete results tables

[link to the paper]

Hanse Hedge
free information sharing for quants and algorithmic traders

https://hanse-hedge.fund

Marius Ciepluch
corrections, replications and disagreement are welcome

marius.ciepluch@gmail.com

DISCLAIMER · REPLACE BEFORE PUBLICATION

This document is published for information and discussion only. It is not financial advice, not an offer or solicitation, and not a recommendation to trade any instrument or adopt any position-sizing rule.

Do not use this document to build or operate a live trading system without independent verification. The results come from simulations and historical windows, with the limits listed on page 23. Past behaviour does not predict future behaviour, and leveraged trading can lose more than the capital committed.

You are responsible for your own testing, risk limits and regulatory position.