

FIELD NOTES FOR THE CFO · DEPLOYMENT STRATEGY

A Hundred Eyes, One Percent

Why a human approval can get weaker as the system it checks gets better, and what airports do instead: plant fake threats and count who finds them

---

Marcos Stipicich

Deployment Strategist · 13 years in supply chain and operations

Cover: Diego Velázquez, Mercury and Argus, about 1659

The argument in one page

An approval only controls what the approver would catch. When errors are rare, people can miss more of them, and the log looks the same.

A person reviews every purchase order and payment the system proposes. The log proves the review happened, not that a wrong one would have been caught. A wrong payment or PO that clears this queue is a loss the signature will not explain to your auditors. If that approval is why the system went live, the deployment rests on a control with an unknown catch rate.

1. **In a lab, the rarer the target, the more people missed it:** 7% at one in two, 30% at one in a hundred.
2. **An approval queue shares one feature with the lab's task:** almost every item is fine. The better the system, the rarer the error.
3. **So test the control, not the signature,** as airports test their screeners.

> Argus had his head encircled with a hundred eyes. Two of them used to take rest in their turns, the rest watched, and used to keep on duty.
> *Ovid, Metamorphoses, Book I, translated by Henry T. Riley, 1851; 1893 edition*

Almost nothing is wrong

On the cover, the end of that watch: Argus asleep. Just outside the crop, Velázquez painted Mercury creeping up on his knees.

Picture a Monday approval queue. A system has drafted the week's purchase orders and proposed the payment run, and someone in finance signs them off. Say it is right ninety-nine times in a hundred. That is a good system, and a strange job: most of the morning nothing is wrong, and the one wrong payment looks like the other ninety-nine.

Where a three-way match already compares the order, the receipt and the invoice, the approver sees only the mismatches, and errors there are common. This piece is about the queue where the approver is the only check.

Argus had a hundred eyes and rested two at a time. The approver has two.

The usual answer is a person in the loop, told to use judgment: the four-eyes principle, applied to a machine. That fails least when errors are common. The better the system, the rarer they are.

Judgment alone is a weak plan when the error is rare.

What the lab found

In 2005, Jeremy Wolfe and colleagues sat twelve paid volunteers in front of a screen and asked them to find “tools” among other objects on a noisy background. It was a made-up version of baggage screening. The variable of interest was how often a tool appeared.

TOOL PRESENT ON	MISSED
50% of trials	7%
10% of trials	16%
1% of trials	30%

Wolfe, Horowitz and Kenner, *Nature*, 2005. At 1%, each volunteer saw twenty targets.

A point system rewarded finding the tool, and false alarms stayed near zero. Misses still rose about fourfold.

The authors’ explanation is mundane. When a target almost never shows up, “nothing here” is almost always right, so people say it faster. At 1%, observers gave up the search in less time than it takes, on average, to find a target. The search simply got shorter.

An approval queue shares one feature with that task: almost every item is fine. Carrying the result over is my bridge: the authors did not know whether the effect appears in the field.

What airports do about it

Screening machines can mix in fake threats. In threat image projection, the machine drops a prerecorded image of a prohibited item into the X-ray picture of a real bag, at random. If the screener finds it, the machine says so. If not, it says that too, and records the miss.

It does not make real threats more common. It measures, on the job, whether this screener is still finding them. In finance terms, it turns a signed control into a tested one.

Three moves, from my reading of the evidence.

1. **Know the rate.** Count how often the system is actually wrong. At one in a hundred, the approver works in the lab's 1% regime, and "stay vigilant" is a wish, not a control.
2. **Test the approver.** Mix known-bad proposals into the queue at a rate you choose and track who catches them. In a study of 1.2 million events at one European airport, about 100 per screener gave a reliability of 0.7 or better, but only when the fakes were hard enough that screeners missed more than one in ten.
3. **Move to a rule whatever a rule can check.** If a check can confirm the number, let it. Keep the human for what the check cannot judge.

The three moves, on Monday

WHAT TO MEASURE	WHAT TO DO
How often approved items turn out wrong	Recheck 300 approved items each quarter. This measures the error the approver let through. At one in a hundred, that is about three errors: a rough rate, not a precise one
How many seeded known-bad items each queue catches: errors the approver is there for, not ones a rule stops	Seed 1 known-bad item per 50 for one quarter, inside the approval workflow, before any document number or ledger entry exists; score the queue, not the person. Pull every fake before it posts, and tell internal audit first
Time per approval, if the workflow logs it	If it falls as the error rate falls, seed that queue first. A faster yes proves nothing alone
How many of the errors found break a rule a check could test	Move that check in front of the approver

The table is my design, not a tested result.

Why the queue first. At one fake per fifty items, the airport's 100 hard fakes take five thousand approvals per person.

Read the result both ways. If the catch rate falls as real errors get rarer, the lab has reached your queue. If it holds, the bridge was wrong.

If it falls, have the approver key in one figure from the source document, with the system's figure hidden, instead of clicking yes. That fix is my design; the sources did not test it.

What this does not prove

- **Fake threats do not cure rarity.** When common targets (44% of trials) and rare ones (10% and 1%) were mixed, a target appeared on 55% of trials, yet the 1% targets were still missed 52% of the time.
- **Easy fakes cannot tell people apart.** A 2005 study of 333 screeners with the standard library found very high hit rates and month-to-month scores too weakly correlated to judge individuals (r = .32 to .58).

One question for your next audit committee. How often is the system wrong, and how would anyone know if the approver stopped looking?

About the author

Marcos Stipicich
Deployment Strategist · Supply Chain & Operations

I grew up on a sheep ranch in Patagonia, where my father's rule was simple: reading is worth nothing if you can't see it on a working day. I still work by it: I measure before I cut, and no tool or method beats what works on the floor.

Thirteen years in Chile, Argentina and Australia taught me that supply chain is not just logistics: it is the whole pipeline, which is why it ends up putting out fires that start elsewhere. A promotion planned in marketing becomes expedited orders in purchasing and a blocked aisle in the warehouse. I would rather find where the leak starts. On one gemba walk, mixed, short and damaged orders led back to a per-operator productivity KPI tied to a bonus; I retired the metric and measured customer satisfaction instead.

I have worked most of that pipeline: planning for more than 200 stores, a three-site logistics operation, a USD 30M cross-border e-commerce P&L and a 50,000-SKU warehouse-system cutover, where I stayed on the floor until the shortcuts died and order processing time fell 20%. Deployments live or die inside that pipeline, and that is where I work. I learned to code in C; today I build with AI coding agents, keeping the math and the rollback in my hands.

Available for pre-cutover forensic audits, deployment strategy, and Forward Deployed roles.

From my manuscript, *The Deployment Chasm*.

linkedin.com/in/marcos-stipicich · marcos.stipicic@gmail.com

Sources

- Buser, D., A. Schwaninger, V. Rehor, and Y. Sterchi. 2025. “Reliability and validity of threat image projection data as a measure of performance in X-ray baggage screening.” *Transportation Research Part A: Policy and Practice* 200: 104640. doi.org/10.1016/j.tra.2025.104640
- Hofer, F., and A. Schwaninger. 2005. “Using threat image projection data for assessing individual screener performance.” *WIT Transactions on The Built Environment* 82: 417–426. doi.org/10.2495/SAFE050411
- Wolfe, J. M., T. S. Horowitz, and N. M. Kenner. 2005. “Rare items often missed in visual searches.” *Nature* 435(7041): 439–440. doi.org/10.1038/435439a

Image. Public domain: Diego Velázquez, *Mercury and Argus*, about 1659. Museo Nacional del Prado (P001175), via Wikimedia Commons.