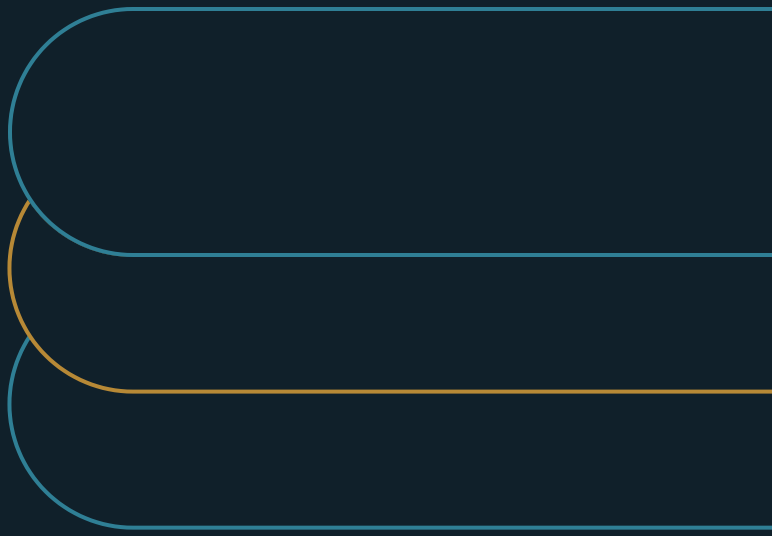
14

WHITE PAPER · PORTFOLIO & DELIVERY GOVERNANCE

Separate the Decision From the Outcome

Why reaching the funded target does not settle whether the outcome—or the decision—was sound



Marco Policani

Enterprise Portfolio · PMO · AI Operating Governance

policani.net · linkedin.com/in/marcpolicani

August 2026

EXECUTIVE SUMMARY

Count the whole outcome, preserve the route, and grade the decision from what was knowable at approval.

A drunk driver can arrive home uninjured and in record time while the trip fails every wider test: the destination was reached, but the decision was reckless, the execution caused harm, and the full outcome was not successful. Portfolio reviews repeat that error when a funded target becomes the whole outcome and the result is read backward into the decision. This paper keeps the full outcome, the execution path, and the approval-time judgment distinct so the organization learns what to repeat and what to repair.

01 A funded target is not the whole outcome: count material effects on cost, risk, capability, customers, and the rest of the enterprise.

02 Preserve the approval-time frame, alternatives, information, ranges, accepted risks, authority, and commitment before the result can edit the record.

03 Keep two grades plus one attribution record: full outcome and decision quality, followed by a reasoned attribution using the execution comparison and variance pattern.

04 Review successful work as well as failures so a favorable headline does not turn a reckless route or weak choice into precedent.

CONTENTS

Executive summary

Suppose a drunk driver arrives home uninjured and in record time. If the scorecard records only his arrival time and condition, the trip looks successful. But he hit a cat, broke several laws, and drove through someone's yard on the

way. The destination was reached. The decision was reckless, the execution caused harm, and the full outcome was not a success.

Portfolio governance makes a quieter version of that error when a funded benefit or an early finish becomes proof of success and good judgment. The narrow target is treated as the whole outcome; the route disappears; then the result is read backwards into the decision. That review rewards luck, punishes disciplined risk-taking, and teaches sponsors to choose short horizons, pad estimates, suppress dependencies, and hide risks that could later be used against them.

A useful retrospective therefore keeps three questions distinct: What was achieved, including collateral effects? How was it achieved? Was the original decision reasonable given what was knowable at approval? This paper turns those questions into two grades and one attribution record. Grade the approval-time decision against a fixed standard. Score the full outcome. Then use the approved-versus-executed route and the variance pattern to attribute any gap to declared uncertainty, execution divergence, or model error. The method covers all four combinations of decision quality and outcome, including the one that does the most durable damage — the weak decision that happened to win and got written into the playbook.

The boundaries matter as much as the method. Observed outcomes ultimately calibrate whether an organization's estimating model is any good. Sound process buys better odds, not results. And a sustained run of misses in the same direction is information about the model, not a reason to keep invoking bad luck.

Panel. A review that records only whether the destination was reached will reward reckless routes, punish disciplined risk, and teach sponsors to serve the scorecard instead of the portfolio.

The narrow target is not the whole outcome

The driving example is deliberately blunt because it exposes a definition that portfolio language often hides. “Outcome” cannot mean only the measure named in the funding paper. Arrival time was one result of the trip, but it was not the whole trip. In the same way, a program can hit its benefit number while bypassing controls, exhausting a critical team, shifting risk into operations, creating platform debt, or damaging customers. Calling that outcome successful because one funded target turned green mistakes the scorecard for the system it is meant to describe.

The distinction does not create three competing scorecards. The review needs two grades and one attribution record. Grade the decision from approval-time evidence. Score the full outcome. Then use the execution comparison — whether the approved route was followed, where controls were bypassed, and which costs or risks were transferred — together with the variance pattern to attribute the gap. The attribution record explains how the result was produced without allowing it to rewrite what was knowable when the choice was made.

Without the separation, a favorable headline can legitimize a reckless method, while an unfavorable headline can discredit a reasonable decision that encountered uncertainty it had honestly priced. Both errors teach the organization the wrong lesson. One turns luck into precedent. The other makes candor and legitimate risk-taking professionally expensive.

Two programs, one inference

Consider two programs at the same quarterly portfolio review. In this composite, the first consolidated four regional order-management systems onto a single platform and closed a quarter early against its benefits case. The second rebuilt the customer identity layer, ran eleven months long, and delivered a fraction of the benefit it was funded to produce. The committee treats the first program's funded target as the whole outcome and approves its sponsor's next proposal in under ten minutes. The second sponsor leaves with a request for further analysis and an informal note that the unit's next request will face a higher bar.

Look at what the two programs had in common. Both were approved through the same stage gate, by the same investment committee, on business cases built to the same template. Both named their dependencies. The consolidation cleared early in part because a regulatory change that would have added a compliance workstream was deferred by twelve months, a decision made outside the company by people who did not know the program existed. The identity rebuild lost most of its schedule when its authentication vendor was acquired and its integration roadmap was frozen for two quarters.

Neither of those events was in either sponsor's control. Neither was in the business case, because neither was foreseeable in the specific form it took. Yet the review treated one as a demonstration of capability and the other as a demonstration of its absence. Nothing in that ninety minutes examined a decision. It examined two results and inferred the decisions backwards from them.

The committee also taught the room something. It taught sponsors that the way to survive a portfolio review is to propose work with short benefit horizons, few external dependencies, and estimates padded far enough that the result cannot embarrass anyone. That is a rational response to the incentive, and it degrades the portfolio.

The measurement error has a name and a literature

This pattern was isolated in the lab almost forty years ago. Baron and Hershey presented people with decisions under uncertainty — a physician weighing whether to operate, with the probabilities stated — and varied only the outcome. Same patient, same odds, same information available at the time. Participants rated the decision itself as better when the patient lived and worse when the patient died, and rated the decider's competence the same way [1]. The information that would have justified a

different rating did not exist at the moment of choice, and the raters knew it. They rated on outcome anyway.

A 2023 high-powered replication of Baron and Hershey's first experiment analyzed 692 participants and recovered the effect at substantial size in both conditions tested — larger where the decision was attributed to a physician than where it was attributed to the patient [2]. The effect is therefore not confined to the original 1988 study.

Field settings show the same shape where the data are good enough to check. Van Ours examined manager replacement in top European football leagues, where performance is measured continuously and publicly. Clubs replaced managers on the basis of recent match results rather than the underlying performance indicators that better predicted future results, and the replacements were, on the evidence, economically inefficient [4]. Football is worth citing here as an empirical setting with unusually clean data on decisions and consequences, not as an analogy about leadership. The transferable point is narrow and uncomfortable: even where decision-makers had strong underlying indicators available, recent results dominated the judgment. Enterprise portfolios often have fewer comparable observations and slower feedback, so they should not assume immunity from the same error.

Mauboussin and Callahan put the practitioner version plainly: where luck materially affects results, the outcome is not a complete causal account of the process that produced it, and the improvable thing is the process [6]. The work is in operationalizing it inside a governance cycle that has to allocate real money in three weeks.

What a decision actually is

Governance cannot grade something it has not defined. A portfolio decision is a commitment made at a specific timestamp, from a specific information set, choosing among a specific set of alternatives, against declared trade-offs, by a named authority. Change any one of those and it is a different decision.

The Society of Decision Professionals sets out six requirements for a quality decision: an appropriate frame, creative and doable alternatives, meaningful and reliable information, clear values and trade-offs, logically correct reasoning, and commitment to action [5]. Treat that as an inspection checklist, not as a causal claim. Nothing in it promises that a decision meeting all six will succeed. What it does is make the decision inspectable — six places to look, each of which can be evidenced or found missing from the record.

The sixth requirement earns its place in portfolio work specifically. A decision that is approved but not resourced, or approved with the losing alternatives still alive in the organization, is not a commitment. It is a deferred argument. Much of what later gets attributed to execution failure is a commitment that was never made.

The hardest of the six to inspect is information, because hindsight contaminates it. Flyvbjerg's outside view supplies the discipline that makes "what was knowable" a checkable question rather than an argument. Where a reference class of comparable prior projects exists, its distribution of cost, schedule, and benefit outcomes was available at decision time, whether or not anyone consulted it [3]. That distinction is the useful one. A base rate that existed and was not consulted is a decision-quality failure. A specific event that no reference class would have predicted — a named vendor being acquired in a particular quarter — is not.

The four cells, and which one costs the most

Cross decision quality with the full outcome and you get four cases. The execution comparison does not add a third axis; together with the variance pattern, it informs why the program entered its cell and where the corrective action belongs. Governance handles two cells adequately by accident and mishandles the other two systematically.

EXHIBIT 1

Use the two grades—and the route—to choose the response

Sound / good Codify the reasoning and stated range, not every tactic that happened to be present.	**Sound / bad** Protect the sponsor; reprice the risk or update the model indicated by the miss.
Weak / good Audit the method before a favorable result turns an untested choice into precedent.	**Weak / bad** Name the missing input or reasoning defect, then repair the relevant control.

Sound decision, good outcome. The easy cell, and the one where the mistake is subtle. The instinct is to codify what the program did. Codify the reasoning instead — the frame, the alternatives considered, the estimate and its stated range. Codifying the tactics of a single successful program is how organizations end up with mandatory methods whose only credential is that they were present when something worked.

Sound decision, bad outcome. The cell that governance destroys. The identity rebuild above lives here if its record holds up: it named its vendor dependency, priced the risk, and lost anyway. Handled well, this cell produces a specific update — the vendor concentration risk was priced too low, and here is the new number — and no change in the sponsor's standing. Handled badly, it produces a sponsor

who will never again put a genuine risk in a business case, and a portfolio that may stop attempting wider-distribution work, including work with high expected value.

Weak decision, good outcome. The expensive cell, because nothing in the organization is motivated to open it. The program landed. The sponsor is promoted. The approach becomes precedent, and the next three programs are pointed at a method that has never actually been tested. A committee that only ever examines failures has, by construction, no way to detect this. Reviewing a sample of successful programs at the same depth as failed ones is a direct mechanism for finding it before the method becomes precedent.

Weak decision, bad outcome. The obvious cell, and still commonly mishandled — because the review names a person rather than a defect. "Sponsor overcommitted" names a person, not a defect. "The business case considered one alternative, and the range presented to the committee was a single point estimate with no reference class" names the defect, and it maps to a fix.

Exhibit 1 — A win and a loss can carry the same decision grade.

Two-by-two: decision quality (graded from the record at the timestamp) on the vertical axis, outcome score on the horizontal. Each quadrant carries the governance response rather than a label: codify the reasoning (sound/good), protect the sponsor, reprice the risk (sound/bad), audit the method before it becomes precedent (weak/good), name the missing input, not the person (weak/bad). Programs from the last four review cycles are plotted as points, which shows the committee at a glance how much of its portfolio it has been grading on the horizontal axis alone.

The decision-quality review

Four moves, run in order. The order is doing work: each move is contaminated if the one before it has not closed.

EXHIBIT 2

Close the retrospective in four ordered moves

01	02	03	04
Reconstruct	**Grade**	**Score**	**Attribute & update**
Rebuild only the frame, alternatives, evidence, ranges, and risks known at approval.	Score the decision inputs and reasoning against a fixed, cited inspection standard.	Count the funded target, collateral effects, and transferred risk on a full-outcome scale.	Classify the gap and change the model, execution control, or accountability record.

Move 1 — Reconstruct the decision at its timestamp

Rebuild the information set as it stood on the approval date. What was in the committee pack. What the paper said was unknown. What range was presented, and with what confidence. Which alternatives were on the table and which had already been eliminated before the paper was written — that second question often finds more than the first.

One rule governs the reconstruction: nothing that arrived after the timestamp enters it. Anything learned since goes in a separate, clearly marked column, because it is the input to Move 4 and poison to Move 2.

Reconstruction is expensive and unreliable when performed after the fact, which is why the practical version is to write the record at approval rather than at review. One page, signed at the gate, filed with the funding decision. The first year's records may be thin and awkward. That thinness is itself a finding about the existing decision process.

Exhibit 2 — Write the decision down before the outcome exists.

One-page decision record, completed at approval and locked: decision statement and date; deciding authority; the frame, including what was explicitly out of scope; alternatives considered and why each was set aside; the two or three assumptions the case depends on; the reference class used and how the estimate compares to it; the stated range for cost, schedule, and benefit, with the confidence attached; the named risks that were accepted rather than mitigated; and what would have to be observed for the decision to be judged wrong. Half a page for anything under the commitment threshold.

Move 2 — Grade the decision against a fixed standard

Take the six requirements [5] and grade each as evidenced, thin, or absent, citing the record. Add one check from the outside view: was a reference class identified, and was the estimate reconciled against it, or was the case built entirely from the inside [3]?

The grade attaches to the decision, not to the decider. That distinction is what makes the grade usable. A sponsor who can lose a grade without losing standing will hand over the record. A sponsor who cannot will produce a record built for the review.

Two failure patterns show up repeatedly in this move. The first is a decision paper with one real alternative and two written to lose. The second is a stated range so wide that no outcome could fall outside it, which reads as candor and functions as immunity. Both are process defects visible from the record alone, independent of what happened next.

Move 3 — Score the full outcome on its own scale

Start with the funded target, but do not stop there. Score benefit realized against the case, schedule and cost against the approved envelope, and the second-order effects the case did not price — the platform debt taken on, the capability built or lost, the customers affected, and material risk shifted into another part of the enterprise. Then inspect how the result was produced: whether controls were bypassed, dependencies ignored, or the approved route materially changed. The outcome score records the complete consequence; the execution comparison supplies evidence for the attribution. Neither is allowed to borrow credit from the decision-quality grade.

Then state what the outcome is evidence of. A single outcome from a decision with a wide declared range is weak evidence about that decision and moderate evidence about the estimate. It is strong evidence about what actually happened: the benefit did not arrive, the money is spent, a customer was harmed, or a risk was shifted. Those facts survive every process argument, and the review should say so plainly before it says anything else.

The rule that makes the separation real: the outcome score never edits the decision-quality grade. The two grades sit side by side, followed by the attribution record, and all three are reported to the committee.

Move 4 — Attribute the gap, then update

Where outcome and expectation diverge, sort the divergence into three buckets: variance inside the declared band, execution divergence from what was decided, or model error. Then update whichever system the attribution points at. Accountability follows the attribution rather than the result.

Uncertainty, execution failure, or a broken model

Three tests, applied in order. None is decisive alone.

Test one: was the result inside the band the decision declared? If the paper stated benefit realization at eighteen to thirty-six months with material dependency risk on a single vendor, and the benefit arrived at thirty-three months after that vendor was acquired, the result remains inside a band the decision had drawn honestly. That is variance, and the correct update is to the vendor-concentration assumption, not to the sponsor. If no band was declared at all, the analysis stops here with a different finding: the decision claimed more certainty than the situation supported, and that is a process defect regardless of what happened.

Test two: did the organization do what it decided to do? Compare the plan as executed against the plan as approved. Scope added without a new decision, sequencing changed after the funding call, a named dependency left unmanaged, a resourcing commitment quietly withdrawn — these are execution divergences, and they belong to the delivery system rather than to the decision. The comparison is only possible if Move 1 produced a record specific enough to diverge from. Vague approvals cannot be violated, which is one reason they persist.

Test three: does the error repeat, and does it repeat with the same sign? One miss may be noise. Nine benefit cases in a row overshooting realized benefit point to a model or execution system that needs investigation. Sign matters more than magnitude here: scattered errors around the estimate suggest an estimating process that may be roughly centered but imprecise, while consistent one-directional error suggests systematic optimism that individual diligence alone will not correct. Reference-class comparison makes this checkable, because it converts a stack of individual judgments into a distribution [3]. This test is also the one that prevents the method from becoming an excuse. Bad luck does not account for a sustained run of losses in the same direction; that pattern is strong evidence that the model, the execution system, or both need correction.

The boundaries between the three buckets are not tidy, and a method that pretends otherwise will be gamed within two cycles. Most execution failures were partly foreseeable at decision time — if the approved plan assumed delivery capacity the organization did not have, the failure belongs to the decision as much as to the delivery. Model error hides inside variance until enough decisions accumulate to expose it, which means early attributions will be revised. Treat the classification as a recorded judgment with stated reasons, revisable when later evidence contradicts it. Recording the reasoning is what makes the revision possible.

Exhibit 3 — Attribute the gap before assigning the credit.

Triage flow from a single program's outcome variance to one of three destinations. Left branch, variance: result inside or near the declared range, no divergence from the approved plan; update the assumption, not the accountability. Centre branch, execution divergence: approved plan and executed plan differ materially; route to delivery governance with the specific divergence named. Right branch,

model error: the same error appears with the same sign across the reference class; route to the estimating standard and the gate criteria. Each branch carries the evidence required to send a program down it, so the classification is arguable rather than asserted.

What changes in practice

Post-mortems. The current form runs backwards from the headline result and produces a narrative in which the outcome was always going to happen. Replace it with two grades and an attribution record: a decision-quality grade drawn only from the timestamped record, a full-outcome score that includes material collateral effects, and the execution comparison that explains how the approved route changed. Write the decision grade before the outcome score is opened. Ban outcome language from the decision section — no "should have known," no "in hindsight" — and require the reviewer to cite the record for every finding. The change is procedural and small, and it substantially alters what these reviews produce.

Benefits-realization reviews. This is the highest-exposure event in the governance calendar, because it arrives with a number that looks objective and invites a single causal story. Run the three-bucket attribution before any reallocation follows from the number. Expect a recurring finding that has nothing to do with luck: benefit cases written in terms no one can falsify. If the case never specified what would count as the benefit not arriving, the review cannot attribute the variance at all, and the correct finding is about the decision standard rather than about the program.

Sponsor evaluation. The football evidence is directly on point here [4]. Where decision-makers had continuous underlying performance data available, recent results still dominated the replacement decision, and the pattern did not pay. Enterprises replace sponsors and reassign programs on much thinner evidence. The practical move is to evaluate sponsors on the decision records they produce — range discipline, alternatives genuinely considered, risks declared in advance rather than surfaced in the post-mortem, commitments honored — accumulated across several decisions, with outcomes entering as the calibration input described below rather than as the grade itself.

Portfolio reallocation. Reallocation is a forward decision made from backward evidence, which makes it a place where outcome bias can do material operating damage. Last cycle's realized benefit is an incomplete basis for next cycle's expected value, particularly for work with long horizons and heavy external dependency. Rank candidate investments on expected value under a stated range, and let decision-quality history inform the confidence attached to the sponsoring unit's estimates rather than the ranking of the investment. If a unit's estimates have run 30 percent optimistic for three years, adjust its cases before comparison, not after approval.

Exhibit 4 — Calibration accumulates across decisions, not inside one.

Running plot, one point per closed decision, showing the signed gap between estimated and realized benefit against the range declared at approval. Points inside the declared band in one color, outside in

another. Read across a sponsoring unit or a reference class rather than down a single program. With enough comparable decisions, a cloud centered near zero is consistent with roughly balanced errors; a cloud sitting persistently on one side is a signal to test and correct the model.

How the method fails

Four failure modes are predictable enough to design against.

EXHIBIT 3

The evidence supports separation, not certainty

Laboratory evidence
The original experiment and direct replication show outcomes alter quality ratings.

Field evidence
Football data show recent results can dominate indicators that better predict performance.

Estimation discipline
The outside view makes available reference-class evidence part of the timestamped record.

Practitioner frame
Decision Quality supplies six inspectable elements without promising a good result.

Not established
No universal attribution rule, causal return, sponsor ranking, or immunity from hindsight.

Process theater. A checklist that generates documents no one would change a decision over. The test is direct: in the last four cycles, has a decision-quality grade changed a funding call, a gate outcome, or an estimating standard? If not, the method is overhead and should be cut or fixed rather than defended.

The luck shield. "Sound process, bad outcome" becomes the sentence that ends inquiry. Two constraints prevent it. First, the sound/bad classification requires the record to have stated in advance what would count as the decision being wrong — a claim made after the loss does not qualify. Second, the classification is counted, by sponsor and by reference class, and a unit whose losses are always attributed to variance has an attribution problem rather than a luck problem.

Hindsight in the reconstruction. Unavoidable if the record is written after the outcome is known, and largely solved if it is written at approval. Where legacy decisions must be reconstructed retrospectively, mark the reconstruction as such and weight its findings accordingly.

Weight without proportion. A one-page record on a small commitment is administration. Set a threshold, apply the full method above it, and use a short form below it. The threshold is a real operating decision with a real cost, and it should be reviewed once a year rather than defended forever.

One more risk sits outside the method: the reviewer's independence. A decision-quality grade written by someone in the sponsor's reporting line will drift upward, and everyone in the room will know it has. Where the portfolio function cannot supply an independent reviewer, rotating reviewers across units is a workable second-best.

The operating move

Start where the record is cheapest to create, which is at the front of the next decision rather than behind the last one.

At the next stage gate above the commitment threshold, require the one-page decision record in Exhibit 2, completed and locked at approval. Do not attempt to backfill the portfolio; retrospective records carry the contamination the method exists to remove. At the next portfolio review, ask three questions in order: What was achieved when the whole consequence is counted? How was it achieved compared with the approved route? Was the original decision sound given what was knowable? Record two grades — decision quality and full outcome — plus one attribution record, and require the committee to state whether it accepts declared uncertainty, execution divergence, or model error before it reallocates anything. Review one successful program at the same depth as the failures each cycle, chosen at random rather than nominated.

After four cycles the portfolio will have its first calibration read: which units declare ranges and which claim certainty, whether estimate errors scatter around zero or sit on one side of it, and how many of the last year's losses were bought at the decision rather than at delivery. That read is the asset. Individual results will keep arriving whether or not anyone grades the decisions that produced them, and a green target will keep being mistaken for a successful trip. The only way to know what the organization should repeat is to count the whole outcome, preserve the route, and write down the judgment before the result can edit it.

Notes and sources

[1] Jonathan Baron and John C. Hershey, Outcome bias in decision evaluation, Journal of Personality and Social Psychology 54(4), April 1988, 569-579. https://pubmed.ncbi.nlm.nih.gov/3367280/

[2] Sriraj Aiyer, Hoi Ching Kam, Ka Yuk Ng, Nathaniel A. Young, Jiaxin Shi, and Gilad Feldman, Outcomes Affect Evaluations of Decision Quality, International Review of Social Psychology 36(1), article 12, July 28, 2023. https://rips-irsp.com/articles/10.5334/irsp.751

[3] Bent Flyvbjerg, Quality Control and Due Diligence in Project Management: Getting Decisions Right by Taking the Outside View, arXiv:1302.2544, 2013. https://arxiv.org/abs/1302.2544

[4] Jan C. van Ours, Outcome bias in managerial decisions, Journal of Economic Psychology 112, January 2026, article 102872. https://www.sciencedirect.com/science/article/pii/S0167487025000844

[5] Society of Decision Professionals, Decision Quality framework / Decision Education at GM. https://www.decisionprofessionals.com/assets/storage/docs/Decision-education-at-GM.pdf

[6] Michael J. Mauboussin and Dan Callahan, Outcome Bias and the Interpreter: How Our Minds Confuse Skill and Luck, Credit Suisse Global Financial Strategies, October 15, 2013. https://www.newedgewealth.com/wp-content/uploads/document-805810220.pdf

About the author

Marco Policani is an enterprise portfolio, PMO, and AI operating-governance leader. He builds portfolio governance systems where AI carries the analytical load and named humans own every decision — an approach documented across case studies, walkthroughs, and working governance guides on his portfolio site.

Portfolio & governance library: policani.net/governance · Full portfolio: policani.net · LinkedIn: linkedin.com/in/marcpolicani