Financial Machine Learning in the Front Office: From Prediction Fit to Portfolio Permission

A practitioner framework for converting machine learning signals into governed portfolio inputs

Renato Guerrieri - Guerrieri Capital Ltd, July 2026

Abstract

Financial machine learning research usually begins by asking whether a model improves forecast error, information coefficient, ranking accuracy or out-of-sample performance. Investment management requires a further decision: whether a particular output should be allowed to influence a portfolio. This paper develops a practitioner framework for that decision. Permission attaches to the intended use, the model version and the portfolio, and can be withdrawn. It is not an automatic consequence of research fit.

The framework distinguishes alpha ranking, sizing, risk filtering, hedge adjustment, execution control and regime monitoring. These applications have different objectives, error costs, implementation paths and monitoring requirements. A model that is unsuitable as primary alpha may still be useful as a secondary signal, risk filter, sizing input, cost forecaster, hedge input, execution policy or monitoring alert. Model class alone therefore does not establish investment relevance.

The paper sets out a signal permission ladder spanning research, validation, frozen shadow deployment, bounded authorisation, restriction, reclassification and termination. The required evidence extends beyond predictive performance to data integrity at the decision date, inference that reflects dependence, exposure decomposition, neutralisation, turnover stability, transaction costs, liquidity, capacity, attribution, model drift, retraining discipline, version control, operational resilience and kill rules agreed in advance. Linear models remain important controls and diagnostic checks. Tree ensembles can be practical front office tools. Deep learning, transformers and reinforcement learning can justify their added complexity in suitable applications, but they carry a higher permission burden.

The central conclusion is direct. Better prediction fit begins the front-office review; it does not complete it. Financial machine learning becomes investable only when a defined output, from a frozen version, survives the full portfolio, implementation and control process for its intended use.

Keywords: financial machine learning; systematic investing; portfolio construction; signal validation; model governance; transaction costs; capacity; attribution; model drift; front-office implementation.

JEL classification: G11; G12; G14; G17; C45; C52; C58.

Executive summary

Financial machine learning is often framed as a contest among model classes. The practical investment decision is different. Capital is not allocated to an architecture. A particular output from a particular model version is permitted to influence a defined portfolio action, within stated limits, because the evidence supports that action and the organisation can observe and stop it.

That distinction raises the standard of review. A model may improve forecast loss while adding little after portfolio construction. It may produce a credible information coefficient but duplicate exposures already present in the book. It may rank securities well but trade too quickly to survive spreads and impact. It may appear differentiated before neutralisation and conventional after beta, sector, size, momentum, liquidity or crowding controls. It may perform under one retraining convention and fail under another. It may be stable in research but fragile when a source is delayed, revised or redefined.

The paper advances eight propositions.

1. **Financial machine learning is not one problem.** Alpha ranking, sizing, risk filtering, hedge adjustment, execution and regime monitoring have different loss functions and consequences of error. Evidence sufficient for an advisory alert may be inadequate for direct position sizing.
2. **Prediction fit is not capital authorisation.** Out-of-sample performance, information coefficients and t-statistics are necessary inputs to review. They do not establish marginal portfolio value, implementation survival, capacity, operational readiness or continued suitability.
3. **The research pipeline can overfit even when the estimator is disciplined.** Search enters through universe definition, timestamps, policy for missing data, feature engineering, labels, horizons, validation windows, neutralisation, hyperparameters, portfolio mapping and cost assumptions. The evidence must be assessed in the context of the full search path.
4. **Model class is not the investment thesis.** Linear models, tree ensembles, neural networks, transformers and reinforcement learning are approximation or policy technologies. The thesis concerns the information source, economic channel, expected horizon and conditions of failure. An architecture name is not an explanation.
5. **Greater representational power normally increases the burden of proof.** Deep architectures may be justified for text, sequences, large predictor panels and heterogeneous inputs. Reinforcement learning may be useful for constrained sequential decisions. Their incremental contribution must, however, compensate for harder attribution, greater sensitivity to training choices and more demanding monitoring.
6. **Linear models remain essential controls.** They provide baselines, decompositions, residualisation, exposure maps and a language for challenge. They often reveal that apparent complexity is repackaging a familiar factor, a data artefact or a portfolio construction effect.

7. **A model can be useful without being primary alpha.** A blocked alpha candidate may deserve a narrower role as a risk filter, exclusion rule, sizing modifier, liquidity forecast, hedge input, execution control or deterioration alert. Reclassification is a valid research outcome, not a lesser result.

8. **Authorisation is conditional and revocable.** It attaches to a defined output, version, portfolio and purpose. Material changes create a successor object and reset the relevant validation clock. Continued authority depends on observable behaviour, clear ownership, bounded actions and credible kill rules.

The proposed permission ladder includes research candidate, validated research signal, frozen shadow model, authorised portfolio input, restricted, blocked and killed states. Authorised portfolio inputs are classified by function rather than placed on a single hierarchy. Ranking, sizing, filtering, hedging and execution each receive separate limits. A model can be approved for one function while remaining blocked for another.

A complete decision record should answer five questions. What portfolio decision will the output influence? What evidence supports that specific role? Which exposures, costs and capacity constraints are embedded in the resulting positions? How will realised behaviour be attributed and monitored? Which events require investigation, restriction, rollback or termination? Until those questions have explicit answers, the model remains a research object.

1. Introduction: why prediction fit is not portfolio permission

Financial machine learning research asks whether a model predicts better. Investment management asks whether the prediction can be sized, traded, monitored, attributed and stopped. The first question concerns statistical learning and generalisation. The second concerns authority over capital.

A raw forecast never enters a portfolio untouched. It is ranked or calibrated, standardised, neutralised, combined with other information, mapped into target positions, constrained by risk and liquidity, translated into trades and exposed to execution. Each transformation can remove, amplify or redirect the apparent research edge. A strong score may add little after construction. A modest predictor may have material value because it controls loss concentration, improves trade selection or changes the timing of risk.

This paper uses **portfolio permission** to describe the decision that a specified output may influence a specified portfolio process under stated limits. The concept is narrower than a general claim that a model is valid and broader than an approval of code. It joins evidence, intended use, capital consequence, implementation, ownership, monitoring and withdrawal conditions. Permission is not permanent. It is tied to a version and can be narrowed, reclassified or removed.

The distinction matters because financial machine learning is not a contest among model classes. Better fit can be economically irrelevant, damaging after implementation or too unstable to govern. An estimator with lower loss can still create a worse portfolio if its incremental information is concentrated in illiquid names, its rank turns over around a hard threshold, its exposures change after retraining or its apparent independence disappears when measured against the existing book. Conversely, a model that would not justify primary alpha authority can improve the process through a bounded, asymmetric role.

The objective is not to oppose complexity. Research has made a persuasive case that nonlinear interactions, large predictor sets and flexible representations can matter in asset pricing and financial forecasting (Gu, Kelly and Xiu, 2020; Kelly, Malamud and Zhou, 2024; Kelly and Xiu, 2023). The objective is to specify what must follow the research result before capital use. Model validation and capital authorisation overlap, but they are not the same decision.

Exhibit 1. Research fit versus portfolio permission

Dimension	Research question	Portfolio permission question
Objective	Does the model predict the stated label out of sample?	Does the output improve the intended portfolio decision after construction and implementation?
Evidence unit	Forecast errors, scores, ranks or labels	Positions, trades, exposures, costs and realised contribution
Comparator	Statistical or machine learning benchmark	Existing portfolio, simple policy and realistic implementation alternative
Dependence	Time, security and label dependence	Dependence plus common trades, constraints, crowding and liquidity
Economics	Plausible source of predictability	Distinct, monetisable contribution after known exposures
Implementation	Often represented by a stylised mapping	Turnover, spread, impact, borrow, financing, latency and capacity
Change	New estimate or retrained model	Potentially a new authorised object requiring fresh review
Monitoring	Forecast performance and data quality	Forecast, exposure, position, trade, attribution, drift and operating behaviour
Failure response	Refit, reject or continue research	Investigate, restrict, reclassify, roll back, block or kill
Decision	Is the result credible enough to study further?	May this version influence this portfolio in this way?

Caption: Research fit establishes predictive credibility. Portfolio permission establishes bounded authority over a defined investment decision.

The remainder of the paper develops this distinction. It first states the contribution and limits of the framework, then places the argument within the financial machine learning literature. It examines model classes through the burden created by their intended use, develops a signal permission ladder, and sets out the evidence needed on dependence, exposure, costs, capacity, attribution, drift and

shadow deployment. Detailed control specifications are placed in the appendices so that the main text remains an investment paper rather than an operating manual.

2. Contribution and Scope

This is not an empirical machine learning performance paper. It reports no proprietary backtest, estimated information coefficient, realised portfolio result or comparison intended to establish that one model class dominates another. The absence of such evidence is deliberate. The paper addresses the institutional decision that follows a research finding rather than presenting a new forecasting contest.

The contribution is a front-office permission framework for financial machine learning. The framework explains how outputs can be classified, shadowed, authorised, monitored, restricted, reclassified or killed. Its unit of analysis is not “the model” in the abstract. It is a particular output from a particular version, proposed for a particular portfolio function. This definition prevents a favourable result in one application from becoming an informal licence for broader use.

The paper does not claim that linear models, tree ensembles, neural networks, transformers or reinforcement learning occupy a universal performance ranking. Model choice depends on data structure, target, horizon, decision form, error asymmetry and operational context. A simpler method can be preferable because it is stable and sufficient. A more complex method can be preferable because it represents information that the simpler method cannot. The framework asks whether the incremental value survives the additional burden created by that choice.

No proprietary models, live signals, holdings, ranks, implementation code, current recommendations or investable products are disclosed. The examples are functional categories, not disguised descriptions of a live process. No evidence is invented to fill those omissions. The paper therefore offers a decision architecture, not a claim of realised performance.

The intended audience is portfolio managers, chief investment officers, allocator research teams, front-office quantitative researchers, systematic equity teams, quantamental teams and risk leads. Each group sees a different part of the same problem. Researchers establish evidence. Portfolio managers define the investment role and capital limits. Risk owners examine exposure, concentration and failure modes. Operations and technology ensure that the approved object can be reproduced and observed. Allocators assess whether the process converts research into controlled investment decisions rather than accumulating models without accountable use.

The framework extends the research discussion around financial machine learning into the portfolio permission layer. It respects the literature on complexity, validation, data snooping, implementation and model risk. It adds a practitioner question: what must be true before an output may alter ranking, sizing,

risk, hedging, trading or monitoring? The answer depends on the use, the model version and the portfolio, and any approval can be withdrawn.

Three boundaries are important. First, the framework is not a substitute for mandate, legal, regulatory, compliance or fiduciary requirements. Second, it does not promise that approved models will perform. Permission is a decision under uncertainty, not a guarantee. Third, it is not a demand for uniform control burden. Control intensity should rise with capital consequence, model complexity, reversibility and the difficulty of detecting failure. A monitoring alert and an execution policy should not pass through identical gates.

The framework is also intentionally symmetrical. It defines routes into use and routes out. A model can be promoted, but it can also be narrowed. An alpha signal can become a risk filter. A sizing input can become an advisory alert. A version can be killed while the underlying research hypothesis remains open. This symmetry is central to disciplined adoption because it allows useful information to survive without preserving authority that the evidence no longer supports.

3. What the financial ML literature gets right

The research foundation for financial machine learning is substantial. The relevant point is not that every published result will survive every implementation. It is that several claims have been established strongly enough to shape serious investment research.

First, nonlinearities and interactions can matter. Asset-pricing predictors are rarely independent, uniformly scaled or stable across states. Flexible methods can capture threshold effects and conditional relationships that a fixed linear specification misses. Gu, Kelly and Xiu (2020) show in a large empirical asset-pricing design that machine learning methods can improve return prediction and portfolio outcomes relative to conventional baselines, with tree-based methods and neural networks among the stronger approaches in their setting. The result supports disciplined investigation of complexity; it does not create a universal model ranking.

Second, large predictor sets can contain useful structure if estimation and validation are controlled. Shrinkage, trees and representation learning provide different ways to manage distributed, correlated and interacting information. Kelly, Malamud and Zhou (2024) make a formal and empirical case for complexity in return prediction. The practitioner implication is not to maximise complexity. It is to compare the incremental contribution of complexity with its estimation, implementation and monitoring cost.

Third, financial machine learning is wider than algorithm selection. Kelly and Xiu (2023) organise the field across prediction, dimensionality reduction, text, asset pricing and portfolio applications. That

breadth supports a central premise of this paper: financial ML is not one problem. The appropriate objective, evidence and control set depend on whether the model is ranking assets, estimating risk, changing exposure, forecasting cost or selecting an action.

Fourth, the structure of financial data makes ordinary validation inadequate. Time dependence, overlapping labels, correlation across the cross-section, revisions and repeated research choices weaken naive inference from random splits. López de Prado (2018) brought wider practitioner attention to data available at the time, leakage, cross-validation design and backtest selection. Bailey, Borwein, López de Prado and Zhu (2017) formalise backtest overfitting as a selection problem, which is the correct frame when many strategies or configurations compete for attention.

Fifth, data snooping and multiple testing are central concerns. White (2000) develops a reality check approach to whether apparent predictive superiority survives the search process. Harvey, Liu and Zhu (2016) show how the proliferation of tested return predictors raises the evidential threshold for new findings. Their concern extends naturally to feature libraries, labels, horizons, architectures, neutralisation choices and portfolio mappings. The effective number of trials is larger than the number recorded in a final experiment table.

Sixth, implementation is part of the economics. Perold (1988) separates paper performance from the cost of obtaining a position. Almgren and Chriss (2001) formalise execution as a trade-off between impact and risk. Novy-Marx and Velikov (2016) and Frazzini, Israel and Moskowitz (2018) show why trading costs and trade design can change the practical value of return signals. Flexible prediction does not repeal spread, impact, financing, borrow or capacity.

The literature therefore supplies the essential research ingredients: complexity may be useful; validation must match financial data; search must be acknowledged; and implementation can dominate gross results. The missing layer is not another general argument for or against machine learning. It is the institutional bridge between credible research and controlled portfolio influence.

4. The front-office gap in financial machine learning

A model can be credible in research and still fail the investment decision. The gap arises because research evidence and portfolio consequences are measured on different objects.

Research commonly evaluates forecasts against labels. The investment process acts on positions and trades. Between the two sit score transformation, neutralisation, blending, constraints, risk scaling, liquidity filters and execution. A result can weaken at any stage. A rank that looks stable across the full universe may be unstable near the portfolio boundary. A forecast with broad accuracy across the cross-section may earn most of its apparent value in names that cannot carry meaningful capital. A model can

add standalone return yet contribute nothing to an existing book because its positions duplicate current exposures.

The first gap is **economic identity**. A complex score may be a nonlinear repackaging of beta, sector, size, value, momentum, quality, low volatility, liquidity, reversal or crowding. That does not make it useless. It changes what it is. A model that times a factor, conditions exposure or anticipates liquidity can be valuable, but it should not be described or governed as independent alpha without evidence.

The second gap is **implementation survival**. Machine learning often discovers local variation, thresholds and interactions. Those features can concentrate turnover or direct the model towards names with high spread, low depth or limited borrow. Gross predictive value is not an economic result until the actual target positions and order distribution survive realistic costs and capacity assumptions.

The third gap is **stability of the authorised object**. A rolling model is not a single static rule. Training data, feature coverage, parameters and representations change. A retraining convention can alter exposures, turnover and failure behaviour even when the architecture is unchanged. Calling every retrain “maintenance” obscures the possibility that the portfolio is receiving a succession of materially different signals.

The fourth gap is **attribution**. A portfolio manager needs to know whether realised contribution came from the intended information, a known factor, a constraint interaction, market timing or favourable execution. A model can remain statistically active while its portfolio contribution migrates. Without stored versions and records for each transformation, that migration may be impossible to diagnose.

The fifth gap is **control feasibility**. Some failures are easy to observe. A missing input, failed job or score outside the approved range can trigger an immediate block. Others are gradual: representation drift, increasing tail concentration, exposure migration or calibration decay. Permission should depend partly on whether deterioration can be observed at the frequency and resolution needed to act.

Exhibit 2. Financial machine learning use map

Use	Decision affected	Primary benefit sought	Dominant failure risk	Typical authority
Alpha rank	Relative security preference	Independent expected return information	Hidden factor or untradeable tail	Direct, but constrained by position limits
Sizing	Position magnitude	Better calibration of conviction, risk or uncertainty	Concentration and nonlinear loss amplification	Direct within hard limits
Risk filter	Reduction or exclusion	Avoidance of adverse states or weak data	Excessive false positives and missed upside	Reductive, usually bounded
Hedge input	Beta, factor or overlay adjustment	Protection against states affecting the whole portfolio	Basis risk, timing error and cost	Direct at aggregate level
Execution control	Order timing, participation or sequencing	Lower implementation shortfall	Simulator error and unstable live response	Direct trading authority within limits

Use	Decision affected	Primary benefit sought	Dominant failure risk	Typical authority
Monitoring	Alert, diagnosis or escalation	Earlier detection of deterioration	Alert fatigue and signals that do not lead to action	Advisory unless separately approved

Caption: The same predictive technology can serve materially different investment functions. Evidence and authority should attach to the function, not to the model label.

The gap can be summarised as a change in the unit of decision. Research asks whether a method generated credible evidence. Portfolio review asks whether a frozen, reproducible output can improve a defined decision after all transformations and can be governed through failure. The first is necessary. The second is the condition for capital use.

5. Model class is not the investment thesis

A model class approximates a mapping or learns a policy. It does not explain why an investment effect should exist. Describing an output as linear, XGBoost, neural, based on transformers or reinforcement learning says something about the machinery and almost nothing about the economic proposition.

The thesis should identify the information source, the route by which it may affect prices or trading outcomes, the likely horizon and the states in which the effect should weaken. It may be behavioural, institutional, informational or related to market structure. It may concern delayed information processing, financing constraints, forced trading, benchmark effects, inventory risk, attention, accounting interpretation or execution frictions. A complete structural model is not required. A proposition that can be challenged is.

This separation improves failure analysis. When architecture and thesis are conflated, a deteriorating model invites a larger model rather than a review of the information source. A tree ensemble may be retuned when the true change is universe composition. A text model may be expanded when document timing has changed. A reinforcement learning policy may be retrained when the reward omitted a cost. The organisation then optimises the mechanism of estimation without reassessing the investment premise.

Model choice still requires a rationale. The method should fit the data geometry and decision form. Penalised linear models may be suitable where many inputs, shrinkage and stability dominate. Trees may be suitable for thresholds, missingness and interactions. Neural networks may be suitable where representation learning adds clear value. Transformers may be justified for text, long sequences or attention across heterogeneous inputs. Reinforcement learning may be justified when the object is a constrained sequence of actions rather than a single forecast.

The rationale must also state why a simpler comparator is insufficient. A gain in a headline forecast statistic is not enough if it disappears after neutralisation, arises from a narrow sample, increases turnover disproportionately or cannot be attributed in the combined portfolio. Incremental value should be assessed at the score, neutralised score, constructed portfolio, portfolio after implementation costs and operating burden.

Exhibit 3. Model class and permission burden

Model class	Useful properties	Characteristic risks	Default burden	Plausible initial role
Linear	Clear decomposition, direct exposure mapping, stable benchmark	Omitted interactions, specification error	Baseline	Control, decomposition, constrained signal
Penalised linear	Shrinkage across many inputs, distributed effects	Penalty sensitivity, unstable selection among correlated features	Baseline to moderate	Secondary signal or blend component
Tree ensemble	Thresholds, interactions, mixed inputs, practical retraining	Split instability, weak extrapolation, misleading importance measures	Moderate	Ranking, filter, sizing or cost forecast under limits
Neural network	Flexible representation of relationships across many inputs	Training path sensitivity, calibration and difficult attribution	High	Frozen shadow use, bounded blend or specialised feature
Transformer	Contextual text and sequence representation	Corpus, tokenisation, embedding and representation drift	High to very high	Feature extraction, secondary signal or monitoring input
Reinforcement learning	Sequential policy optimisation under constraints	Simulator learning, reward misspecification and feedback instability	Very high	Execution, turnover, hedge or sizing policy within hard bounds

Caption: Burden rises with the number of ways an output can be convincing in research yet unsuitable in portfolio use. Capital consequence can outweigh architectural complexity.

The table states defaults, not a universal ranking. A linear model that determines gross exposure can require more scrutiny than a transformer that produces an advisory text feature. A carefully bounded neural model can be less dangerous than a poorly specified regression embedded directly in sizing. The relevant rule is to approve the decision pathway, not the architecture. Model class determines some controls; intended authority determines the rest.

6. Linear models as controls, not enemies

Linear models remain central to a serious machine learning process because they make structure visible. Their value does not depend on believing that financial relationships are linear. They serve as benchmarks, decompositions, residualisation tools and diagnostic controls.

The first role is comparison. A complex candidate should be tested against a sensible linear or penalised linear baseline using the same inputs available at the time, labels, universe, windows and portfolio mapping. This separates gains from model flexibility from gains created elsewhere in the pipeline. If both models improve after a data repair, that improvement belongs to the data process. If the nonlinear

model wins only before costs, its flexibility may be capturing variation too short-lived or too expensive to trade.

The second role is exposure decomposition. Projecting scores and positions on familiar characteristics provides a first description of what the candidate is doing. It is not a causal proof. It identifies whether beta, sector, region, size, value, momentum, quality, low volatility, liquidity, volatility, reversal or crowding explain a material part of the output. The residual can then be evaluated separately from the explained component.

The third role is sanity checking. Linear relationships can expose sign errors, scaling problems, timestamp mistakes and dependence on a small set of variables. A complex model may show attractive aggregate performance while simple controls reveal unstable signs or implausible sensitivity to revised fields. That finding does not automatically reject the model. It identifies where the burden of proof should rise.

The fourth role is monitoring. Rolling projections of a nonlinear score can reveal migration in beta, liquidity or style exposure. A materially different loading after retraining may indicate a new portfolio object even when the research team describes the update as routine. Linear approximations also provide a compact language for explaining where the model disagrees with the benchmark and where those disagreements are concentrated.

Feature importance is not a substitute for portfolio decomposition. Importance measures can distribute credit across correlated variables, depend on the chosen method and say little about the contribution of the resulting trades. Breiman's distinction between data modelling and algorithmic modelling remains useful: predictive performance and explanatory structure are different objectives, and both matter in investment review (Breiman, 2001b).

A model that loses its apparent novelty under linear controls may still have value. It may improve factor timing, risk scaling, trade selection or nonlinear implementation. The correct response is classification, not defence of the model label. Linear models are therefore part of the control environment that makes complex methods usable.

7. Tree ensembles as front-office operating tools

Tree ensembles occupy a practical middle ground. They represent thresholds and interactions without requiring the full training and monitoring apparatus of deeper architectures. Random forests reduce the variance of individual trees through aggregation (Breiman, 2001a). Gradient boosting builds an additive sequence of trees against a chosen objective, while XGBoost supplies an efficient regularised

implementation (Chen and Guestrin, 2016). In empirical asset pricing work, tree-based methods have often been competitive where conditional relationships matter (Gu, Kelly and Xiu, 2020).

Several properties support their practical use. Trees accommodate mixed distributions and nonlinear response shapes without requiring every transformation to be specified in advance. They can identify that a variable matters only beyond a threshold or in combination with another characteristic. Training and rescoring are generally manageable across multiple windows, seeds and challenger designs. A fixed data snapshot, feature specification and parameter file can usually reproduce the model without a large representation pipeline.

These advantages do not make trees transparent. Split structures can change sharply when correlated variables substitute for one another. Importance can be dispersed across related features or concentrated because of scale, cardinality or missingness. A model may fit historical thresholds with no stable economic meaning. Piecewise constant predictions can behave poorly outside observed support. Hyperparameter search can select depth, learning rate, subsampling and regularisation choices that happen to favour the validation design.

Missingness deserves particular attention. Convenient handling of missing values can cause the model to learn coverage, vendor or timing patterns rather than the intended information. A change in data availability can then move predictions even when the economic state has not changed. Tests should isolate missingness by source, date and universe segment rather than treating it only as a technical feature.

A permission case for a tree ensemble is strongest when the nonlinear gain remains after portfolio mapping and neutralisation; rank, exposure and turnover are stable across reasonable perturbations; the result is not confined to a narrow sample or data convention; the training object is reproducible; and the intended role does not demand explanatory precision the model cannot supply.

Tree ensembles can be particularly useful as secondary alpha, risk filters, sizing modifiers and implementation forecasts. In those roles, conditional structure can add value without granting one model sole authority over direction or risk. They are not a default answer. Their practical attraction is that they often provide substantial nonlinear capacity at an operating burden that an investment team can challenge, version and monitor.

8. Deep learning and transformers: power, opacity and monitoring burden

Deep learning becomes relevant when the information set is too large, heterogeneous or unstructured for a fixed feature design to represent efficiently. Text, long sequences, event histories and large

predictor panels can justify models that learn intermediate representations rather than receiving only economically labelled variables.

The case is clearest for text. Filings, transcripts, announcements and news contain information in wording, context, sequence and interaction. Treating them as data requires choices about document availability, revisions, deduplication, time zones, parsing, tokenisation, entity resolution, context length and aggregation. Gentzkow, Kelly and Taddy (2019) describe the economic and statistical issues involved in text analysis. Ke, Kelly and Xiu (2019) examine return prediction using text data. Attention architectures introduced by Vaswani et al. (2017) broaden the capacity to represent context and long-range relationships.

Deep models can also combine many weak inputs, learn security or state representations and join cross-section and time series structure. A representation may be valuable even when the final portfolio rule remains simple. For example, a text or sequence model can produce a bounded feature that is consumed by a separately controlled ranking or risk process. Separating representation from capital action can reduce the authority given to the most complex component.

The higher burden begins with data lineage. A text model is not defined only by weights and architecture. It is defined by the exact corpus that was available at each decision time, the extraction and cleaning rules, the model used for tokenisation or embedding, and the treatment of later corrections. A historical archive reconstructed from current documents can contain information that was unavailable or differently formatted in real time. A corpus can also drift because filing practices, source coverage, language or document length change.

Training path is another source of variation. Initialisation, batching, optimiser settings, stopping rules, architecture choice and source used before training can alter the representation. Hyperparameter search can consume a large implicit trial budget. The winning configuration may reflect validation noise, particularly when labels are weak and samples are dependent. Repeated tuning against the same period turns that period into training evidence regardless of how it is labelled.

Interpretability is partial. Attention weights, saliency, gradients, ablations and local attribution can be useful diagnostic tools, but none provides a complete economic explanation. They can show sensitivity under a chosen representation without establishing causality or portfolio contribution. A PM still needs to know which exposures, names, events and trades carry the result. The relevant attribution chain runs from raw input through representation and score to target position, constrained position, order and realised outcome.

Representation drift deserves separate monitoring from forecast decay. The score distribution can remain stable while the internal embedding of sectors, language or states changes materially.

Conversely, embedding statistics may move without harming the intended portfolio function. Monitoring should therefore connect representation diagnostics to downstream effects: rank stability, exposure, turnover, calibration, concentration and marginal contribution.

Deep models also increase operational dependency. A change in a library, model checkpoint, numerical setting, hardware path or upstream parser can change output. Reproducibility requires pinned versions, immutable training data, recorded seeds and a complete specification of preprocessing. A candidate that cannot be recreated from controlled artefacts cannot receive meaningful authority for that version.

The initial permission should usually be narrower than the eventual ambition. A deep model may begin as a feature generator, secondary signal or monitoring input. Broader influence should follow only when its incremental contribution survives simpler comparators, exposure and cost tests, frozen shadow deployment and review across versions. The issue is not opacity in itself. It is whether opacity prevents the organisation from understanding capital consequences and detecting deterioration.

9. Reinforcement learning: policy value and simulator risk

Reinforcement learning differs from supervised prediction because it learns a policy over states and actions. The object is not only an estimate of a future label. It is a rule for choosing actions in sequence to maximise a reward under an environment (Sutton and Barto, 2018). Financial applications include execution, portfolio allocation and hedging (Hambly, Xu and Yang, 2023).

That distinction is useful because many investment decisions are sequential. An execution policy trades off impact, urgency and residual risk over time. A sizing rule responds to changing conviction, volatility and cost. A hedge policy adjusts protection as exposures and market states evolve. A turnover policy decides when not to trade. These are policy problems rather than isolated forecasts.

The principal danger is that an agent learns the simulator rather than the market. The environment necessarily simplifies price formation, liquidity, impact, delay, fills, competitor response, borrow, financing and market states. An agent can exploit these simplifications in ways that appear optimal in simulation and fail immediately outside it. The richer the action space and the longer the horizon, the more opportunities exist for reward exploitation and compounding error.

Reward design creates a second failure channel. A policy optimises what is specified, not what was intended. A return objective without realistic costs invites excessive turnover. A risk-adjusted reward may hide tail concentration or path dependence. A penalty can create corner solutions. A terminal objective can encourage behaviour that looks sensible only at the end of an episode. If a desired behaviour is important, it should not be assumed to emerge from an indirect reward when it can be imposed as a hard constraint.

Off-policy evaluation is also difficult in markets. Historical data reveal outcomes under actions that were actually taken, not under every alternative action. Estimated counterfactuals depend on assumptions about market response and state coverage. Policy performance can therefore look precise while relying on weak support in the most important regions. Stress tests should focus on states where the proposed action differs most from a simple policy and where historical coverage is least convincing.

For these reasons, reinforcement learning is more defensible when the action space is bounded, the objective is operationally observable and a deterministic fallback exists. Execution, turnover control, hedge adjustment and constrained sizing often meet those conditions better than unconstrained raw alpha discovery. The policy can be compared with a simple schedule or rule, limited by participation and exposure caps, and disabled without reconstructing the whole investment thesis.

Permission should cover the complete policy system: state variables, environment, action set, reward, constraints, training method, exploration assumptions and fallback. An environment change is a model change. A revised impact function, liquidity model or reward weight can create a new policy even if the code structure is unchanged.

Shadowing an RL policy requires more than comparing notional performance. Review should examine proposed actions, state coverage, constraint use, disagreement with baseline policy, sensitivity to environment assumptions, tail sequences and operational exceptions. The team should know not only whether the policy would have improved an average result, but how it behaves when liquidity withdraws, prices gap, data fail or the state lies outside the training support.

Reinforcement learning is therefore a policy technology with a high burden of proof. Its use is strongest when sequential action genuinely matters and authority can be bounded. It is weakest when a flexible simulator is treated as evidence that an unconstrained agent has discovered a durable market law.

10. The ML signal permission ladder

The permission ladder provides a common vocabulary for research, shadowing, authorisation and withdrawal. It is not a ranking of model quality and not a linear promotion path. It classifies what an output may do now, on which version, for which portfolio and under which limits.

Exhibit 4. ML signal permission ladder

State	Meaning	Permitted activity	Exit condition
Research candidate	Hypothesis and implementation under development	Exploratory research in controlled environments	Advance on documented evidence; reject or redesign on failure
Validated research signal	Evidence package has passed independent review for stated claims	Reproducible research and construction tests	Freeze for shadowing or return to research after material change

State	Meaning	Permitted activity	Exit condition
Frozen shadow model	Exact version runs on live data without capital authority	Record scores, positions, trades, costs, exposures and exceptions	Authorise a bounded role, extend shadow, restrict or block
Ranking input	Output may affect relative preference	Influence ranking within approved blend and constraints	Continue, reduce weight, reclassify or withdraw
Sizing input	Output may affect position magnitude	Adjust sizing within hard risk and concentration limits	Continue, cap, revert to baseline sizing or withdraw
Risk filter	Output may reduce, exclude or flag exposure	Apply approved reductive rule with exception handling	Continue, recalibrate, make advisory or withdraw
Hedge or overlay input	Output may alter aggregate exposure	Adjust approved hedge or overlay within a corridor	Continue, narrow corridor, revert or withdraw
Execution or turnover input	Output may influence orders or decisions not to trade	Act within participation, liquidity and fallback controls	Continue, restrict action set, revert or withdraw
Restricted	Authority has been narrowed pending review	Only the stated residual function remains active	Restore after evidence, reclassify, block or kill
Blocked	Output has no portfolio authority	Research may continue; no capital influence	Return only through a new evidence and shadow process
Killed	Version is retired and cannot be restored informally	Historical analysis only	A successor is treated as a new object

Caption: Authorisation is functional and reversible. A model may occupy different states for different outputs or portfolios.

The first transition is from research candidate to validated research signal. Validation should confirm that the data, labels, transformations, training, inference and portfolio tests reproduce the claimed evidence. It should also challenge the claim itself: whether the benchmark is appropriate, whether the search path is represented, whether the economic identity is understood and whether the proposed use is narrower or broader than the evidence.

Validation does not confer capital authority. It establishes that the candidate is credible enough to freeze and observe. The shadow version should be immutable. Live inputs are processed on schedule, proposed positions and trades are recorded, and all downstream transformations are reconstructed without sending orders. Any material change to features, target, window, retraining rule, model, neutralisation, portfolio mapping or cost model creates a new shadow object. Otherwise, the shadow period becomes continuing research under a misleading label.

Authorised states are defined by function. Ranking authority allows influence over relative preference, not position magnitude outside the approved construction rule. Sizing authority can alter capital concentration and therefore requires calibration, tail and portfolio-loss evidence beyond rank quality. A risk filter is reductive but can still destroy value through false positives or correlated exclusions. Hedge authority acts at portfolio level and can create basis and timing risk. Execution authority acts directly on orders and requires hard operational limits.

Restricted status is important because failure is not binary. A model may retain value in a narrower role while a concern is investigated. A primary rank may be reduced to a secondary blend component. A sizing model may be prevented from increasing positions but allowed to reduce them. An execution policy may retain a recommendation not to trade while losing control of participation. Restrictions should be explicit and time-bounded, not unapproved workarounds that preserve the original authority.

Blocked and killed are different. Blocked means the output has no current authority, but the hypothesis may remain under research. Killed means the specific version is retired and cannot be reinstated through an operational override. A successor must be identified, validated and shadowed as a new object. This distinction protects against the common tendency to reactivate a failed version after a short period of favourable retrospective results.

Permission is also specific to the portfolio. A signal may fit a diversified long-only portfolio but be unsuitable for a concentrated or market-neutral mandate. Existing exposures, liquidity, capital scale, turnover budget and risk constraints change the marginal result. The same frozen score can therefore receive different authority across portfolios without inconsistency.

The ladder creates discipline by making status observable. Every active output should have an owner, version, authorised function, limit set, monitoring record and current state. Research quality remains central, but the organisational question becomes clear: what exactly is this object allowed to change?

11. IC, t-stat and the illusion of independent evidence

Information coefficients and t-statistics are useful summaries only when the evidence unit and dependence structure are correct. Machine learning can generate a large number of predictions without generating a large number of independent observations.

A daily cross-section may contain thousands of securities, yet common market, sector, macroeconomic and liquidity shocks link their forecast errors and returns. Treating every security-date observation as independent can produce implausibly precise inference. Petersen (2009) shows why dependence across firms and time matters for standard errors in finance panels and compares clustering approaches. The appropriate method depends on the design, but the general requirement is fixed: security and date dependence must be reflected rather than ignored.

Overlapping labels create another problem. A forecast of a multi-period return produced every day reuses much of the same future path. Adjacent observations are not independent trials. Multiple horizons can add apparent breadth while sharing outcomes. Event labels can overlap across securities because the events themselves are driven by common information. Effective sample size should be considered in relation to the economic process, not inferred from the number of rows.

The information coefficient also depends on construction choices. Pearson and rank correlation answer different questions. Cross-sectional and time-series pooling have different interpretations. Winsorisation, weighting, universe filters and missing-value treatment can change the result. A positive average IC can hide concentration in a small period or universe segment. A stable aggregate can coexist with unstable tails, which is where a portfolio may place most of its capital.

Inference must include the search context. A reported result is usually the survivor of choices over features, targets, horizons, architectures, penalties, windows, seeds, neutralisation and portfolio rules. White (2000), Harvey, Liu and Zhu (2016), and Bailey et al. (2017) address complementary parts of this selection problem. Their common implication is that a winning specification should be judged against the process that produced it, not as a single test specified in advance.

This is especially important for a nominal holdout. A period ceases to be independent evidence when researchers repeatedly inspect it and adapt decisions in response. The same is true of a live shadow period if the model is altered after each observation. Labels such as validation, test and shadow do not create independence. Organisational behaviour determines whether evidence has been consumed.

The strongest evidence package uses several units. It examines date-level portfolio outcomes, rank behaviour across dates and securities, security-level error concentration, regime and subperiod stability, and the result of fully mapped positions after costs. It reports the dispersion of outcomes rather than only an average. It compares the candidate with simple baselines and with the existing portfolio. It records unsuccessful branches sufficiently to place the selected result in context.

Statistical significance and economic materiality should also be separated. A small effect estimated precisely may not survive implementation or add to the book. A larger but noisy effect may be useful as a bounded diversifier or risk input. The decision should consider consequence, reversibility and diversification, not convert a t-statistic into a capital weight.

Evidence should finally match the proposed authority. Primary alpha requires persistent incremental return information after controls and costs. Sizing requires calibration and concentration evidence. Risk filters require analysis of false positives, missed adverse states and opportunity cost. Execution requires realised or implementation evidence close to the live process. A statistic is not strong or weak in the abstract. It is sufficient or insufficient for a particular decision.

12. Neutralisation, attribution and hidden exposure tests

A complex model can predict well because it identifies a familiar exposure more efficiently than a simple specification. The correct response is neither automatic rejection nor automatic acceptance. The exposure must be identified, measured and incorporated into the model’s classification.

Neutralisation should be applied at more than one layer. Projecting the raw score on known characteristics shows what the forecast resembles before construction. Rebuilding the portfolio under explicit constraints shows what remains after the optimiser and portfolio rules act. Comparing the candidate with existing signals reveals whether it adds marginal information or merely changes the timing and concentration of exposures already held.

The standard review should include beta, sector, region, size, value, momentum, quality, low volatility, liquidity, volatility, reversal and crowding proxies. The exact definitions should match the investment universe and be set before reviewing the preferred result. A model can appear neutral under a narrow set of factors and remain exposed to an alternative horizon, nonlinear tail or interaction that matters in stress.

Exhibit 5. Hidden exposure and neutralisation matrix

Exposure	Score test	Portfolio test	Stress lens	Possible reclassification if dominant
Beta	Score sensitivity to market beta	Net and gross beta contribution	Sharp rallies, sell-offs and gap events	Beta timing or hedge input
Sector and region	Group means, ranks and tail concentration	Active group weights and risk contribution	Group rotations and regional shocks	Allocation or constraint input
Size	Relation to capitalisation and tradability	Positions by size bucket, turnover and days to trade	Withdrawal of small cap liquidity	Sizing with liquidity constraints
Value	Loading to valuation composites	Factor contribution and tail overlap	Long value drawdowns and reversals	Value timing or blend component
Momentum	Loading across horizons and rank overlap	Momentum contribution and common trades	Momentum crash states	Momentum timing or risk filter
Quality	Relation to profitability and balance-sheet measures	Quality factor and defensive contribution	Recovery and stress transitions	Quality tilt or drawdown filter
Low volatility	Relation to beta and volatility	Defensive concentration and risk contribution	Volatility spikes and rebounds	Sizing with volatility constraints
Liquidity	Relation to spreads, volume and turnover	Participation, spread cost and liquidation time	Withdrawal of depth and volume	Liquidity forecast or exclusion rule
Reversal	Dependence on recent returns	Contrarian contribution and turnover	Gaps, rebounds and forced unwinds	Overlay over short horizons or execution input
Crowding	Relation to ownership, flow or proxies for common positions	Trade overlap and correlated exit risk	Deleveraging and crowded liquidation	Capacity control or warning signal

Caption: Neutralisation determines economic identity. A disappearing residual can support reclassification even when it does not support a claim of independent alpha.

A complex model need not be orthogonal to every known factor. Many investment processes intentionally combine compensated exposures, timing and selection. The requirement is to know what the model adds and which part of the outcome should be expected to behave like a familiar factor. Hidden beta is a problem when it is mistaken for stock selection, not because beta is always forbidden.

Tail analysis matters more than average correlation. Two signals can have low correlation over the full sample and select the same positions during stressed periods. Overlap in the tails of the ranking, analysis of common trades and conditional exposures should complement linear projection. The portfolio should examine where the candidate disagrees with existing signals because those disagreements often determine marginal return and drawdown behaviour.

Attribution must be designed before use. The record should preserve raw inputs, versioned scores, transformed scores, target positions, constrained positions, orders, fills and cost estimates. Realised or shadow contribution can then be divided into intended signal, known exposure, construction, constraint, trading and residual components. Classic attribution provides a useful discipline of separating sources of outcome, although machine-learning portfolios require a more granular chain than traditional policy allocation analysis (Brinson, Hood and Beebower, 1986).

The test of independence is ultimately marginal. A standalone portfolio can look attractive while adding little to an existing book. The relevant comparison introduces the candidate into the actual construction process and measures incremental return, risk, drawdown, turnover, cost, concentration and capacity. It also asks whether the candidate makes the rest of the portfolio more or less stable.

If a model disappears after neutralisation, three conclusions are possible. The first is rejection because the claimed independent effect is absent. The second is reclassification because the model captures a useful factor timing, risk or implementation function. The third is narrower authorisation because only a residual component survives. The evidence decides among them. The principle is economic honesty: the model should be called and controlled according to what it actually does.

13. Turnover, transaction costs, liquidity and capacity

Prediction becomes capital use through trades. The conversion is not a final deduction from an otherwise complete backtest. It is part of the investment hypothesis.

Signal half-life and rebalance frequency must be considered jointly. A short-lived relationship may require trading when spreads are wide or competition is concentrated. A slower schedule can sacrifice gross information while improving net value. The relevant choice is the combination of forecast horizon, rebalance timing, turnover tolerance and execution method, not the forecast horizon in isolation.

Turnover should be analysed as a distribution. Average annual turnover can hide bursts after retraining, index changes, earnings periods, volatility shocks or data transitions. Those bursts can exceed available liquidity and coincide with common systematic trading. Review should examine turnover by date, security, sector, region, liquidity bucket, signal tail and reason for trade.

Rank stability is not enough. Names near a portfolio threshold can enter and leave repeatedly while overall rank correlation remains high. Small score moves can create large orders if sizing is nonlinear or constraints bind. Implementation tests should therefore use target positions and orders, not infer trading from score statistics.

Perold's implementation shortfall distinguishes paper return from the cost of obtaining a position (Perold, 1988). Almgren and Chriss (2001) formalise the trade-off between market impact and price risk. Machine learning does not change the underlying economics. A cost review should include the elements relevant to the mandate, which can include spread, fees, market impact, delay, opportunity cost, borrow, financing and taxes. Assumptions should vary with liquidity, volatility, order size, urgency, direction and market state.

Liquidity should be treated as a distribution and a binding resource. Average daily volume can overstate executable capacity when volume is episodic, driven by events or concentrated at the close. The process should examine participation, days to trade, days to liquidate, concentration in the least liquid positions and the interaction between model conviction and tradability. Stress assumptions should recognise that liquidity is often weakest when risk reduction is most urgent.

Novy-Marx and Velikov (2016) show that trading costs materially affect the implementability of anomaly strategies and that choices used to reduce costs change net outcomes. Frazzini, Israel and Moskowitz (2018) document the relation of costs to trade size and liquidity using large execution data. The practical conclusion is not that strategies with high turnover cannot work. It is that cost survival must be demonstrated on the candidate's actual trade distribution and with a credible implementation process.

Capacity is a curve, not a label. It depends on capital, participation, turnover, concentration, synchronisation, signal decay and execution. A model can be valuable at one scale and have negative marginal value at another. It may support more capital when positions are diversified and patient, or less when edge is concentrated in illiquid tails. Capacity analysis should expose assumptions and show how net contribution changes as scale increases.

The surrounding portfolio matters. A new signal may have moderate standalone turnover but conflict with existing holdings, increasing gross trading. It may trade the same names at the same time as other signals, creating internal and external crowding. It may instead reduce turnover by confirming positions or acting as a filter against trading. The relevant cost and capacity measure is marginal to the combined book.

Crowding adds state dependence. Khandani and Lo (2011) show how common factor exposures and forced unwinds contributed to the August 2007 quant dislocation. A model that overlaps common trades may face widening spreads, correlated exits and reduced depth precisely when risk limits require

action. Decay after publication provides a related warning: documented anomalies can weaken as investors learn and capital responds (McLean and Pontiff, 2016). Machine learning can recombine information; it cannot remove competition or market adaptation.

A candidate does not pass implementation review because a conservative constant cost was subtracted. It passes when the team can explain where and when it trades, how cost varies, how capital scale changes net value, how the signal behaves under liquidity stress and how the model interacts with the existing portfolio. When those answers are weak, authority should be reduced, the construction should change or the model should remain outside capital use.

14. Model drift, retraining and version control

A financial model can deteriorate because the relation it estimates changes, because its inputs change, or because the organisation changes the model itself. These sources of drift require different responses.

Feature drift concerns the distribution, coverage or meaning of inputs. A variable can move because the market changed, because the universe changed, because a vendor revised a definition or because a processing rule was altered. Distributional change is not automatically harmful. It becomes relevant when the model is extrapolating, when missingness patterns move or when downstream exposure changes.

Label drift concerns the target and its economic meaning. Return horizons can change behaviour as execution, competition or market structure evolve. A cost label can become stale when trading style or venue composition changes. A risk event can be defined consistently yet become more or less frequent. Monitoring the target is therefore part of monitoring the model.

Relationship drift concerns the mapping from inputs to outcomes. A predictor may decay, reverse or become dependent on market state. This is the failure most often meant by model decay, but it is difficult to identify cleanly because live samples are short, noisy and dependent. A deterioration alert should lead to diagnosis rather than automatic refitting.

Portfolio drift occurs when the score's economic identity changes. The forecast distribution can look normal while beta, sector, liquidity, turnover or tail concentration migrates. For an authorised investment input, portfolio drift can matter more than small changes in forecast loss because it alters the risk actually being taken.

Retraining is not a neutral maintenance operation. A rolling window changes the sample. Re-estimating changes coefficients, splits or representations. Hyperparameter retuning changes the search. Feature

additions or removals change the information set. A new data source changes lineage. Any of these can create a successor signal even when the model family and output name remain constant.

The practical distinction is between **scheduled retraining within an approved rule** and **material model change**. A scheduled retraining can remain within an existing permission only if the training window, data specification, feature set, target, hyperparameters, randomisation policy, portfolio mapping and acceptance tests are themselves approved and the resulting behaviour remains within defined bounds. A material change resets the relevant validation and shadow clock.

Materiality should be judged by both method and outcome. A small code change can materially alter ranks. A large internal refactor can leave outputs identical. Version comparison should therefore cover score distribution, rank changes, tail membership, factor and liquidity exposures, turnover, cost, capacity, concentration and expected portfolio contribution. The question is whether the capital object changed, not how many lines of code changed.

Fixed and rolling models present different trade-offs. A fixed model offers a stable object and clearer attribution, but can become stale. A rolling model can adapt to changing information, but creates continual version risk and can chase noise. There is no universal answer. The permission record should state why the chosen approach fits the use and how the organisation distinguishes adaptation from instability.

Challenger models help separate deterioration from normal noise. A challenger may use a simpler architecture, a different window or an alternative feature group. It should not automatically replace the incumbent because it recently performed better. Its purpose is to provide diagnostic contrast and a prepared alternative. Replacement remains a new authorisation decision unless the substitution rule was explicitly approved in advance.

Version control must cover more than code. A reproducible model object includes data snapshots or immutable references, feature definitions, preprocessing, training configuration, software dependencies, random seeds where relevant, fitted artefacts, scoring logic, construction rules, cost assumptions and monitoring thresholds. The record should support reconstruction of what the portfolio knew and did at a given time.

The governing principle is that permission attaches to a versioned process, not to a continuing project name. Retraining can produce a new signal. When that possibility is ignored, the portfolio can accumulate unreviewed change under the appearance of ordinary maintenance.

15. Shadow deployment, monitoring and kill rules

Shadow deployment is the bridge between research and capital use. It tests whether a frozen object can run on live data, produce the expected downstream decisions and remain observable under ordinary operating conditions. It is not an extended backtest and not a period for continuous model improvement.

The shadow object must be fixed. Its data policy, features, target, training rule, fitted artefact or scheduled retraining rule, score transformation, portfolio mapping, cost model and monitoring specification should be versioned before the period begins. Live scores, target positions and notional orders should be produced on the same schedule that would apply after authorisation. Data delays, corrections and operational exceptions should be recorded as they occur, not repaired invisibly after the fact.

A shadow period serves four purposes. First, it checks operational reproducibility: whether the pipeline runs, inputs arrive and outputs reconcile. Second, it observes behaviour not visible in historical reconstruction, including current missingness, data latency, rank movement and trade concentration. Third, it creates a clean record for attribution and cost comparison. Fourth, it tests whether the proposed dashboard and escalation process are capable of detecting meaningful change.

Shadow evidence should not be overstated. A short live observation period cannot establish performance over long horizons or behaviour in rare states. It can, however, invalidate a candidate quickly through data failure, unexpected turnover, exposure migration, unstable rankings or inability to reconstruct decisions. The period is therefore both an operational test and a check on the realism of the research representation.

If the model is modified during shadow, the validation clock resets for the affected claims. A minor correction may require only targeted re-review if outputs are demonstrably unchanged. A change in data, features, training, neutralisation, portfolio mapping or costs creates a new object. Otherwise, the team observes a moving target and later attributes favourable results to a version that never existed consistently.

Exhibit 6. Shadow deployment and kill rule loop

Stage	Required object	Decision question	Possible action
Freeze	Versioned model, data policy, construction and thresholds	Is the candidate fully reproducible and its proposed role explicit?	Enter shadow or return to research
Run	Live scores, positions, trades and operational logs	Does the object operate as specified without hidden intervention?	Continue, correct immaterial issue or reset
Observe	Forecast, exposure, turnover, cost,	Is behaviour consistent with the evidence	Continue, extend, investigate or

Stage	Required object	Decision question	Possible action
	capacity and attribution record	and intended economic identity?	restrict proposed role
Review	Independent assessment of evidence and exceptions	Is a bounded portfolio function justified?	Authorise, reclassify, block or extend shadow
Monitor	Live dashboard linked to the version and ownership	Is the authorised function behaving within expected and hard limits?	Continue, investigate, restrict or roll back
Kill or replace	Recorded breach, loss of thesis, control failure or approved successor	Can authority continue without discretionary optimism or informal override?	Kill version; validate and shadow any successor

Caption: Shadowing and monitoring form one control loop. A material change returns the object to an earlier stage rather than preserving authority by name.

Monitoring should distinguish **expected range indicators** from **hard limit controls**. A breach of the expected range may trigger investigation because the model is behaving unusually, while a hard limit breach should prevent or reduce action. Conflating the two creates either excessive shutdowns or thresholds that are ignored because they were never intended to be binding.

Kill rules should be agreed in advance where possible. Relevant triggers include loss of data integrity; inability to reproduce or reconcile output; use of an unapproved version; material exposure or concentration outside limits; implementation cost beyond the approved case; persistent loss of marginal contribution; failure of the economic thesis; operational instability; or a control breach that makes behaviour unobservable. No single universal metric can define model failure. The rule set should combine hard events with a structured judgement process.

Performance alone should rarely be an instantaneous kill trigger unless the loss breaches a portfolio risk limit. Forecasts and strategies experience noise. A drawdown can be consistent with the thesis, while a favourable result can conceal data leakage or hidden beta. The review should examine whether losses arise from expected risk, factor movement, trading, data failure, model drift or a change in the economic mechanism. Kill rules are not a substitute for diagnosis; they ensure diagnosis occurs before additional authority is granted by inertia.

A credible rule also defines the state after a trigger. Authority may be suspended, restricted to a reductive function, rolled back to a prior approved version or removed completely. Ownership and response time should be clear. A model should not remain active merely because responsibility for the decision is diffuse.

Exhibit 7. Front office monitoring dashboard for ML signals

Panel	Core observations	Primary diagnostic question	Escalation path
Data and operations	Timeliness, missingness, revisions,	Did the authorised object receive	Block output on integrity failure;

Panel	Core observations	Primary diagnostic question	Escalation path
	coverage, job failures and reconciliation	and process the intended information?	investigate source or pipeline
Score and rank	Distribution, calibration, rank stability, tail membership and disagreement with baseline	Has the forecast changed in form or location?	Review drift, retraining and support
Exposure and concentration	Beta, factor, sector, region, liquidity, crowding, name and group concentration	Is the model taking the risk it was authorised to take?	Constrain, neutralise, restrict or reclassify
Trading and capacity	Turnover, order concentration, participation, spread, impact, days to trade and capacity use	Can the current output be implemented within the approved case?	Reduce scale, alter construction or suspend trading role
Attribution	Intended signal, known exposure, constraints, trading and residual contribution	Is realised value coming from the claimed source?	Investigate migration; reduce or change authority
Model and version	Current version, retrain status, challenger difference, feature and representation drift	Is this still the approved object?	Reset review for material change; roll back if required
Risk and kill status	Limit use, drawdown context, exceptions, open investigations and kill rule state	Is continued authority justified now?	Continue, restrict, block or kill

Caption: The dashboard links model behaviour to portfolio consequences. Detailed field definitions and review frequencies appear in Appendix C.

Monitoring is useful only when it supports action. A dashboard with many indicators but no ownership, thresholds or escalation route becomes an archive. A smaller set of measures tied to the approved function is preferable to a large catalogue that nobody can interpret. The objective is to identify whether the model remains the object that was authorised, whether it is still useful in that role and whether the portfolio can exit safely if it is not.

16. How ML enters a portfolio

Machine learning can enter a portfolio through several channels. The permission threshold should reflect the channel because each creates a different form of capital consequence.

A **primary alpha rank** has broad influence over relative positions. It requires evidence of incremental return information after neutralisation, construction, costs and capacity. The model should improve the combined portfolio rather than only a standalone sort, and its influence should be capped until live attribution supports expansion.

A **secondary alpha signal** or **blend component** receives less authority but still affects direction. Its value may lie in diversification of information or model error. The review should focus on marginal contribution, tail overlap and behaviour when it disagrees with stronger incumbents. A low average correlation is insufficient if the signals converge in stressed states.

A **risk filter** reduces or blocks exposure under defined conditions. Because it is reductive, it can appear safer than alpha, but false positives can impose a persistent opportunity cost and create concentrated exclusions. The filter should be tested on what it removes, when it removes it and whether the action is reversible. A filter can be approved as advisory before it becomes automatic.

An **exclusion rule** is a stronger version of a filter. It may be appropriate when data integrity, tradability or risk asymmetry makes a false negative more costly than a false positive. Because the action is absolute, ownership, exception handling and reason logging are important. A model should not create unexplained prohibited names.

A **sizing input** can change portfolio risk more than a ranking signal. It may scale positions by forecast strength, uncertainty, liquidity, drawdown risk or state. The evidence must address calibration, concentration, path dependence and the possibility that model confidence is most wrong in unusual states. Hard position, factor, gross and liquidity limits remain outside the model.

A **beta, hedge or overlay input** changes aggregate exposure. Its value should be judged in the states where protection or participation is intended, after hedge cost and basis risk. A model that predicts a market state weakly may still be useful if the action is small, asymmetric and diversifying. The linked action, not the classifier alone, requires approval.

A **drawdown warning** or **regime classifier** can begin as an advisory signal. Its quality depends on timeliness, state stability and the cost of false alarms. A regime label has no investment meaning until paired with an action. Different actions, such as reducing gross, changing hedge or limiting turnover, require separate evidence.

A **cost or liquidity forecaster** affects which trades are taken and how they are sized. It should be calibrated against realised implementation and reviewed for selection effects. A cost model can improve net performance by preventing expensive trades even if it contains no return forecast. Its errors can still bias the portfolio towards particular sectors, sizes or market states.

An **execution policy** chooses timing, sequencing, urgency or participation. It acts directly on orders and therefore requires operational limits, reconciliation and deterministic fallback. Evaluation should include implementation shortfall, fills, opportunity cost, stability across order types and behaviour under stressed liquidity. Policy authority should be narrower than the training action space unless every action has been reviewed.

A **monitoring alert** is advisory and usually has the lowest capital consequence. Its usefulness depends on whether it detects an actionable condition early enough and whether an owner can distinguish a true concern from noise. An alert should not acquire de facto authority through repeated manual action without a separate approval of that action.

The same model may support several channels, but evidence should not transfer automatically. A text representation could support ranking, risk and monitoring. A volatility model could support sizing and hedging. A liquidity model could support exclusion and execution. Each output and action pair should have its own limits and monitoring.

A model blocked as alpha can therefore remain valuable. It may identify a factor state, forecast cost, reduce weak positions or flag crowding. This is not a softening of standards. It is a more precise match between evidence and authority. The objective is not to place every model in the portfolio. It is to use information where it has a defensible function and to withhold it where it does not.

17. Front office governance checklist

A public framework should state the governing questions without turning the main paper into a procedural handbook. Detailed controls appear in Appendix B. At decision level, an investment committee, PM or risk owner should be able to answer the following.

Area	Governing question
Objective	What exact decision will the output influence, and what is explicitly outside its authority?
Thesis	What information or market mechanism supports the expected value, horizon and conditions of failure?
Data	Can the information set available at the time, lineage, revisions and missingness be reconstructed?
Label and evidence	Is the target economically aligned with the proposed action, and are dependence and search reflected in inference?
Model choice	Why is this model class suitable, and what does it add beyond a credible linear or simple policy benchmark?
Exposure	Which known factors, groups, liquidity characteristics and crowded trades explain the score and portfolio?
Implementation	Do turnover, spread, impact, financing, borrow, latency and capacity support the proposed scale?
Portfolio role	Does the candidate add marginal value to the actual book after constraints and existing signals?
Attribution	Can score, position, constraint, trade and realised contribution be reconstructed by version?
Change	Which retraining and maintenance actions are approved in advance, and which changes create a successor object?
Shadow and monitoring	Has a frozen version operated on live data, and can deterioration be detected at the required speed?
Limits and exit	Who owns the output, what limits bind, and what causes restriction, rollback, blocking or termination?

The checklist is deliberately centred on use. A model inventory without a map of actual influence is incomplete. A complex model that generates an unused research score may have little capital consequence. A simple model embedded in gross exposure or execution can have substantial consequence. Governance should follow purpose, authority, reversibility and detectability.

Current banking model-risk guidance offers useful lifecycle principles, although its legal scope differs from the investment organisations addressed here. The 2026 US interagency guidance emphasises a risk-based approach linked to model purpose, use and business risk. The current PRA SS1/23 sets principles covering identification, governance, development, use, independent validation and mitigants (Board of Governors of the Federal Reserve System, Federal Deposit Insurance Corporation and Office of the Comptroller of the Currency, 2026; Prudential Regulation Authority, 2026). The relevant analogy is proportionality and lifecycle control, not direct transplantation of a bank supervisory framework.

Independent review should combine verification and challenge. Verification asks whether the documented process was implemented and reproduced. Challenge asks whether the proposed use follows from the evidence, whether simpler alternatives were dismissed too quickly, whether costs and capacity are credible, whether known exposures explain the result and whether the monitoring process can detect the stated failure modes. Capital authorisation requires both.

Good governance does not mean that every adjustment requires a committee. Approved in advance operating ranges can support efficient action. The essential point is that the ranges, ownership and escalation path are explicit before performance pressure arrives. Informal exceptions should not become permanent extensions of authority.

18. Limitations

This framework does not solve the central uncertainty of investment management. A model can pass every stated control and fail in live use. Out-of-sample evidence is historical evidence, not a guarantee that the future relation will persist.

The framework can also be applied badly. A long checklist can create the appearance of discipline while difficult judgements are avoided. Thresholds can be selected to permit a preferred model. Independence can become formal rather than substantive. Documentation can describe a process that is not followed under pressure. The quality of challenge matters more than the volume of control language.

Machine learning can overfit the full research process. Data available at the time and walk-forward tests do not remove search across universes, features, labels, models, horizons, construction rules and costs.

The effective trial count is difficult to measure, and corrections for selection rely on assumptions. Evidence should therefore be treated as graded rather than converted into a claim of certainty.

Complexity can hide simple exposures. Neutralisation and attribution are themselves dependent on the model and cannot prove economic independence. Factor definitions differ, proxies are incomplete and nonlinear or exposures that depend on market state can remain. Feature importance does not provide a complete economic explanation.

Implementation estimates are uncertain. Market impact, borrow, financing, liquidity and capacity can change with scale and state. Historical cost data may not represent a stressed exit. A model can alter the market response to its own trades as capital increases. Capacity analysis is therefore conditional, not a permanent certification.

Drift is difficult to distinguish from noise in real time. Automatic retraining can chase variation; fixed models can become stale. Challenger comparisons help but do not remove the decision problem. Reinforcement learning adds the specific risk that a policy learns the assumptions and gaps of its environment rather than a transferable market response.

A public paper cannot disclose proprietary data, live signals, holdings, code or internal evidence. The framework is consequently described without empirical case material that would be required for an actual authorisation decision. It is a permission model, not a substitute for the evidence package of a particular portfolio.

Finally, proportionality is dependent on context. Portfolios differ in mandate, liquidity, holding period, capital scale, operational structure and tolerance for model error. The framework should be adapted without weakening its central discipline: authority must match evidence, consequence and the ability to withdraw.

Conclusion

Financial machine learning earns front-office relevance only when it becomes a governed portfolio input. Better prediction fit is the beginning of the discussion, not the end.

A model deserves authority only if a specific output adds something after known exposures, survives realistic trading, fits within liquidity and capacity, and produces contribution that can be attributed. The authorised version must be reproducible. Changes must be classified rather than hidden inside maintenance. Deterioration must be observable. Actions must be bounded. A credible kill rule must exist before performance pressure makes withdrawal difficult.

This standard does not favour simplicity for its own sake. Linear models remain controls and decompositions. Tree ensembles can be practical operating tools. Deep learning and transformers can justify their burden where representation matters. Reinforcement learning can support sequential policy where the environment, action set and fallback are tightly specified. No architecture receives a general licence. The decision is attached to the output, use, portfolio and version.

Nor does the standard require every useful model to become primary alpha. A candidate may be more valuable as a risk filter, sizing modifier, cost forecast, hedge input, execution policy or monitoring alert. Reclassification can preserve useful information while preventing an unsupported claim from acquiring capital authority.

The durable institutional advantage is therefore not only a predictive signal. It is the capacity to distinguish research fit from portfolio permission, to convert evidence into bounded action, and to remove that action when the evidence or operating conditions no longer support it. Financial ML is not a contest among model classes. It is a portfolio process under uncertainty.

Author note

Renato Guerrieri is the principal of Guerrieri Capital Ltd, a private research vehicle focused on systematic investment research and implementation infrastructure. The paper reflects a practitioner view of financial machine learning as a front-office portfolio process, with emphasis on signal validation, implementation permission, portfolio construction, attribution, monitoring and model governance.

Disclosure and disclaimer

This paper is methodological and research-oriented. It describes a front-office framework for evaluating financial machine learning signals and model outputs. It does not disclose proprietary models, live signals, holdings, recommendations, implementation code or investable products.

The paper is provided for research and discussion. It does not constitute investment advice, a recommendation, an offer, a solicitation or a representation that any process will achieve a particular result. Investment decisions require assessment of mandate, risk, liquidity, legal, regulatory, tax, operational and other considerations specific to the investor and portfolio.

AI-assisted drafting disclosure

AI-assisted drafting and editing tools were used in preparing this manuscript. The author remains responsible for the paper’s arguments, structure, review, factual accuracy and final content.

References

Almgren, R. and Chriss, N. (2001) 'Optimal execution of portfolio transactions', *Journal of Risk*, 3(2), pp. 5-39. doi: 10.21314/jor.2001.041.

Bailey, D. H., Borwein, J. M., López de Prado, M. and Zhu, Q. J. (2017) 'The probability of backtest overfitting', *Journal of Computational Finance*, 20(4), pp. 39-69. doi: 10.21314/jcf.2016.322.

Board of Governors of the Federal Reserve System, Federal Deposit Insurance Corporation and Office of the Comptroller of the Currency (2026) *Supervisory Guidance on Model Risk Management*. SR 26-2, 17 April. Available at: https://www.federalreserve.gov/supervisionreg/srletters/SR2602.htm (Accessed: 22 June 2026).

Breiman, L. (2001a) 'Random forests', *Machine Learning*, 45(1), pp. 5-32. doi: 10.1023/a:1010933404324.

Breiman, L. (2001b) 'Statistical modeling: The two cultures', *Statistical Science*, 16(3), pp. 199-231. doi: 10.1214/ss/1009213726.

Brinson, G. P., Hood, L. R. and Beebower, G. L. (1986) 'Determinants of portfolio performance', *Financial Analysts Journal*, 42(4), pp. 39-44. doi: 10.2469/faj.v42.n4.39.

Chen, T. and Guestrin, C. (2016) 'XGBoost: A scalable tree boosting system', in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785-794. doi: 10.1145/2939672.2939785.

Frazzini, A., Israel, R. and Moskowitz, T. J. (2018) 'Trading costs', working paper, 7 April. doi: 10.2139/ssrn.3229719.

Gentzkow, M., Kelly, B. and Taddy, M. (2019) 'Text as data', *Journal of Economic Literature*, 57(3), pp. 535-574. doi: 10.1257/jel.20181020.

Gu, S., Kelly, B. and Xiu, D. (2020) 'Empirical asset pricing via machine learning', *Review of Financial Studies*, 33(5), pp. 2223-2273. doi: 10.1093/rfs/hhaa009.

Hambly, B., Xu, R. and Yang, H. (2023) 'Recent advances in reinforcement learning in finance', *Mathematical Finance*, 33(3), pp. 437-503. doi: 10.1111/mafi.12382.

Harvey, C. R., Liu, Y. and Zhu, H. (2016) '... and the cross-section of expected returns', *Review of Financial Studies*, 29(1), pp. 5-68. doi: 10.1093/rfs/hhv059.

Ke, Z. T., Kelly, B. T. and Xiu, D. (2019) 'Predicting returns with text data', NBER Working Paper 26186. doi: 10.3386/w26186.

Kelly, B. T., Malamud, S. and Zhou, K. (2024) 'The virtue of complexity in return prediction', *Journal of Finance*, 79(1), pp. 459-503. doi: 10.1111/jofi.13298.

Kelly, B. T. and Xiu, D. (2023) 'Financial machine learning', *Foundations and Trends in Finance*, 13(3-4), pp. 205-363. doi: 10.1561/0500000064.

Khandani, A. E. and Lo, A. W. (2011) 'What happened to the quants in August 2007? Evidence from factors and transactions data', *Journal of Financial Markets*, 14(1), pp. 1-46. doi: 10.1016/j.finmar.2010.07.005.

López de Prado, M. (2018) *Advances in Financial Machine Learning*. Hoboken, NJ: Wiley.

McLean, R. D. and Pontiff, J. (2016) 'Does academic research destroy stock return predictability?', *Journal of Finance*, 71(1), pp. 5-32. doi: 10.1111/jofi.12365.

Novy-Marx, R. and Velikov, M. (2016) 'A taxonomy of anomalies and their trading costs', *Review of Financial Studies*, 29(1), pp. 104-147. doi: 10.1093/rfs/hhv063.

Perold, A. F. (1988) 'The implementation shortfall: Paper versus reality', *Journal of Portfolio Management*, 14(3), pp. 4-9. doi: 10.3905/jpm.1988.409150.

Petersen, M. A. (2009) 'Estimating standard errors in finance panel data sets: Comparing approaches', *Review of Financial Studies*, 22(1), pp. 435-480. doi: 10.1093/rfs/hhn053.

Prudential Regulation Authority (2026) *SS1/23: Model Risk Management Principles for Banks*. Supervisory Statement 1/23, current version published 23 April. Available at: https://www.bankofengland.co.uk/prudential-regulation/publication/2023/may/model-risk-management-principles-for-banks-ss (Accessed: 22 June 2026).

Sutton, R. S. and Barto, A. G. (2018) *Reinforcement Learning: An Introduction*. 2nd edn. Cambridge, MA: MIT Press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I. (2017) 'Attention is all you need', *Advances in Neural Information Processing Systems*, 30, pp. 5998-6008.

White, H. (2000) 'A reality check for data snooping', *Econometrica*, 68(5), pp. 1097-1126. doi: 10.1111/1468-0262.00152.

Appendix A. ML signal permission taxonomy

This appendix separates lifecycle status from portfolio function. A model first has a status, such as research, shadow, authorised, restricted, blocked or killed. An authorised model also has one or more defined functions, such as ranking or risk filtering. The two dimensions should not be compressed into a single label.

A.1 Lifecycle states

State	Purpose	Minimum evidence or control	Capital authority	Typical transition
Research candidate	Develop and challenge a hypothesis, data process and model	Named objective; initial thesis; design with time stamps; benchmark; research record	None	Advance to formal validation, redesign or stop
Validated research signal	Establish that the stated research claims and process are credible	Independent reproduction; review of the search context; exposure, cost and capacity analysis; proposed role	None	Freeze for shadowing or return to research after material change
Frozen shadow model	Observe the exact candidate on live data and test operations	Immutable version; scheduled scoring; stored downstream decisions; dashboard; exception log	None	Authorise a bounded function, extend shadow or block
Authorised	Permit one or more defined portfolio functions within limits	Signed decision record; limits specific to the role; owner; monitoring; attribution; kill rules	Only the approved functions	Continue, expand through review, restrict, reclassify or withdraw
Restricted	Preserve a narrower function while a concern is addressed	Written restriction; residual authority; end condition; fallback	Only the stated residual function	Restore, reclassify, block or kill
Blocked	Prevent any portfolio influence while research may continue	System prevention; reason; owner; conditions for return	None	Return only through a new evidence and shadow process
Killed	Retire a version and prevent informal reactivation	Final decision record; archived artefacts; disabled production path	None	Any successor is treated as a new object

A state should have an effective time, decision owner and linked version. “Under review” is not a sufficient status if downstream systems cannot determine whether the output may still act. Where review takes time, the interim state should be normal, restricted or suspended according to the rule agreed in advance.

A.2 Authorised portfolio functions

Function	What the output may change	What it may not change without separate approval	Principal evidence emphasis	Typical external controls
Primary ranking input	Relative security preference across the approved universe	Position size, gross exposure or hedge outside the construction rule	Marginal return information, neutralised contribution, cost survival and capacity	Cap on blend weight, exposure constraints, liquidity rules
Secondary signal or blend component	A bounded share of combined ranking or conviction	Sole control of direction or unconstrained override	Diversification of information, disagreement analysis and tail overlap	Stable blend rule, contribution limit, correlation review
Sizing input	Position magnitude within approved eligibility and direction	Breach of name, group, factor, gross, net or liquidity limits	Calibration, uncertainty, concentration and drawdown behaviour	Hard position and risk limits outside the model
Risk filter	Reduction, deferral or flagging of exposure	Increasing exposure unless separately approved	Cost of false positives, coverage of adverse states and action stability	Maximum reduction, exception process, reason logging
Exclusion rule	Removal of an asset, trade or data state from eligibility	Unexplained discretionary reinstatement	Error asymmetry, data quality and operational clarity	Escalation, exception authority, expiry or review rule
Hedge or overlay input	Aggregate beta, factor, volatility or protection position within a corridor	Changes beyond the approved overlay or to unrelated exposures	Protection by state, basis risk, timing and cost	Exposure corridor, instrument eligibility, fallback hedge
Cost or liquidity forecast	Trade selection, urgency, participation or size adjustment	Return direction unless separately validated	Calibration to realised trading and selection effects	Participation, liquidity, reconciliation and override limits
Execution or turnover control	Order timing, sequencing, participation or decision not to trade	Orders outside approved instruments, venues, sizes or action set	Implementation shortfall, fill quality, operational stability	Hard action limits, deterministic fallback, kill switch
Regime or drawdown input	Advisory state label or approved bounded response	Any action not separately specified	State stability, timeliness, false alarms and action value	Separate action approval, exposure cap, review requirement
Monitoring alert	Notification, diagnosis and escalation	Direct portfolio action	Detection quality, lead time and actionability	Named owner, service expectation, suppression and escalation rules

Authorisation for one function does not imply authorisation for another. A model may be approved to reduce positions while remaining prohibited from increasing them. A liquidity forecast may affect trade timing without affecting expected return. A regime classifier may generate an alert while no automatic portfolio response is permitted.

A.3 Minimum permission record

Every authorised output should have a concise record containing:

Field	Required content
Identity	Stable model, signal and output identifier
Version	Code, data rule, feature, parameter, construction and monitoring versions
Authorised function	Exact decision the output may influence
Scope	Portfolio, universe, horizon, schedule, instruments and eligible actions
Limits	Influence, concentration, exposure, turnover, liquidity, participation or action bounds
Evidence basis	Validation package, shadow period and decision date
Economic identity	Intended source of value and known exposure components
Owners	Research, portfolio, data, technology and independent review responsibilities
Monitoring	Required panels, expected ranges, hard limits and review cadence
Change rule	Retraining approved in advance and changes that require a new review
Fallback	Prior version, simple rule or operating process used when authority is suspended
Exit rule	Conditions for investigation, restriction, blocking and termination
Review status	Effective date, current state and next scheduled review

The record should be understandable without reconstructing the research history. The validation archive can contain the full evidence, but the active permission must state what the portfolio may do now.

A.4 Materiality and transition rules

Materiality should be assessed in two ways. **Method materiality** concerns changes to data, target, features, architecture, hyperparameters, training window, retraining rule, randomisation, score transformation, neutralisation, portfolio construction, cost model or monitoring. **Outcome materiality** concerns changes in ranks, tails, exposure, turnover, concentration, cost, capacity or expected portfolio contribution.

A change can be material under either test. A small code correction that materially alters ranks requires a new review. A substantial internal refactor that produces identical controlled outputs may not. The decision and evidence should be recorded rather than inferred from the developer's description of the change.

The following transition principles apply:

9. A validated candidate enters shadow only after the object and proposed function are frozen.
10. A material change during shadow creates a successor and resets the affected evidence period.

11. Expansion from one authorised function to another requires evidence specific to the role.
12. Expansion in capital, universe or action range requires a capacity and consequence review even if the model is unchanged.
13. Restriction should specify what remains permitted, who can restore authority and the conditions for restoration.
14. Blocking prevents portfolio influence but can allow continued research.
15. Killing retires the version. A later favourable backtest does not reinstate it.
16. A successor may reuse the underlying economic thesis, but it does not inherit the predecessor's permission automatically.

A.5 Portfolio and version specificity

Permission is specific to the portfolio because the surrounding book determines marginal exposure, liquidity, turnover and capacity. The same score can be acceptable in a diversified portfolio and unsuitable in a concentrated sleeve. It can reduce risk in one mandate and duplicate risk in another. A central validation package may support several decisions, but each portfolio should record its own authorised function and limits.

Permission is also specific to each output. One model can generate a return rank, uncertainty estimate and cost forecast. Those outputs may have different evidence quality and different owners. Treating the model as a single approved entity can allow a weak output to inherit the reputation of a stronger one.

Version specificity protects attribution. When historical results, live decisions and monitoring all refer to the same frozen object, the organisation can determine what changed and why. Without that discipline, a project name can conceal a succession of signals, and favourable evidence can be attributed selectively across versions.

Appendix B. Front office validation checklist

The checklist is organised by decision stage. It is not a scoring model. Certain failures, including unresolved future information, unreproducible data, an undefined portfolio function or the absence of a safe fallback for direct authority, can block permission regardless of strengths elsewhere.

B.1 Investment proposition and intended use

Control question	Evidence required	Decision relevance
Is the proposed function explicit?	Primary rank, secondary signal, sizing, filter, hedge, execution, cost or monitoring classification	Sets the relevant burden and prevents authority drift

Control question	Evidence required	Decision relevance
Is the economic thesis stated?	Information source, mechanism, horizon, expected conditions and failure states	Separates model class from investment proposition
Is the consequence of error described?	Potential return, risk, concentration, trading and operational harm	Determines limits, reversibility and review depth
Is a simple alternative identified?	Current rule, linear model, conventional optimiser or baseline policy	Establishes the relevant comparator
Is the benefit measurable?	Incremental return, risk reduction, lower cost, greater capacity, stability or earlier detection	Aligns validation with the intended decision
Are uses that are not authorised explicit?	Functions and portfolios outside the request	Prevents a narrow result from becoming a general licence

B.2 Data, lineage and timing integrity

Control question	Evidence required	Decision relevance
Are sources approved and documented?	Source, licence scope, owner and permitted use	Confirms lawful and operable access
Is availability reconstructed on a point-in-time basis?	Observation, publication, receipt and processing timestamps; revision policy	Excludes future information and unrealistic latency
Are identifiers and universe history reconstructed?	Corporate actions, mappings, delistings, mergers and eligibility by date	Prevents survivor, mapping and universe bias
Are original vintages or defensible reconstructions used?	Revision history and treatment of corrected or restated data	Matches research information to live information
Is missingness understood?	Coverage by time, source, region, sector and security type	Detects unintended learning from coverage patterns
Are vendor and schema changes monitored?	Change notices, field reconciliation and historical impact assessment	Distinguishes market drift from data drift
Can the full pipeline be reproduced?	Immutable snapshots or references, transformations, dependencies and run manifest	Supports validation, attribution and rollback

B.3 Target, sampling and decision alignment

Control question	Evidence required	Decision relevance
Does the target match the proposed action?	Exact outcome, horizon, measurement time and economic interpretation	Prevents optimisation of an irrelevant label
Is outcome overlap mapped?	Shared future windows, sampling interval and event overlap	Defines the effective evidence unit
Are execution assumptions embedded correctly?	Achievable decision time, price, delay, fill and cost treatment	Avoids labels based on unavailable outcomes
Are class or tail definitions fixed?	Threshold logic and frequency through time	Prevents retrospective target selection
Was universe selection point-in-time?	Eligibility, liquidity and mandate rules by date	Prevents survivor and coverage bias
Are sampling and observation weights	Date, security, group or event weighting	Prevents large groups or periods from

Control question	Evidence required	Decision relevance
justified?	rationale	dominating unintentionally

B.4 Model development and search control

Control question	Evidence required	Decision relevance
Is the rationale for the model class documented?	Relationship between data geometry, decision form and chosen method	Confirms that complexity serves the problem
Are linear and simple policy baselines run?	Same inputs, target, windows and portfolio mapping	Isolates incremental value from pipeline improvements
Is feature selection inside the training process?	Selection by window, scaling, imputation and preprocessing	Prevents future information entering feature design
Is hyperparameter tuning separated from final evaluation?	Training, tuning and untouched evaluation boundaries	Protects the evidential meaning of the final period
Is the trial history recorded?	Material features, targets, horizons, models and portfolio variants	Places the selected result in search context
Is randomness controlled and measured?	Seed policy, repeated fits and dispersion of outcomes	Detects dependence on a favourable training path
Are missingness and coverage stressed?	Alternative treatments, missing indicators and tests for source withdrawal	Detects vendor or coverage artefacts
Are ablations performed?	Removal of feature groups, representations and model components	Identifies the source of incremental value
Is calibration tested where magnitude matters?	Predicted versus realised outcomes by score range and time	Determines whether probability or sizing use is justified

B.5 Validation evidence and statistical inference

Control question	Evidence required	Decision relevance
Is evaluation chronological and walk-forward?	Repeated estimation, tuning and testing using only available information	Approximates live reuse rather than one historical split
Is dependence reflected?	Date, security, group, event and treatment of overlapping labels	Prevents false precision
Is stability examined?	Period, universe, region, state, seed and results across reasonable specifications	Distinguishes broad evidence from a narrow winner
Is judgement adjusted for search applied?	Trial inventory, untouched evidence and sensitivity to researcher choice	Reduces confidence driven by selection
Is evidence reported as a distribution?	Means, dispersion, concentration and tail behaviour	Prevents averages from hiding unstable economics
Does the result survive a simple benchmark?	Forecast, portfolio and cost comparison with baseline	Tests whether complexity adds rather than decorates
Is latency represented?	Data arrival, computation time and decision	Confirms that the claimed action is

Control question	Evidence required	Decision relevance
	schedule	feasible
Are failure states coherent with the thesis?	Conditional signs, regimes, horizons and known adverse conditions	Supports diagnosis and monitoring design

B.6 Exposure, neutralisation and attribution

Control question	Evidence required	Decision relevance
Are market and group exposures decomposed?	Beta, sector, region and other group loadings relevant to the mandate	Identifies broad risk embedded in the score and positions
Are style exposures decomposed?	Size, value, momentum, quality, low volatility and volatility	Identifies familiar return sources and concentration
Are implementation exposures decomposed?	Liquidity, spread, turnover, reversal, borrow and crowding proxies	Identifies tradability and risk of common exits
Are raw and constrained objects both tested?	Score projection and reconstructed portfolios with specified neutralisation	Shows whether construction preserves or redirects value
Are tail and stressed overlaps examined?	Top and bottom ranks, common trades and correlation conditional on state	Detects duplication missed by average correlation
Is marginal value measured in the actual book?	Candidate introduced into existing construction and constraints	Establishes incremental rather than standalone contribution
Can attribution be reconstructed?	Raw score, transformed score, targets, constraints, orders, fills and costs	Supports diagnosis and continued authority
Is economic classification honest?	Evidence for alpha, factor timing, risk, cost or policy function	Determines the correct permission state

B.7 Portfolio construction, turnover, cost and capacity

Control question	Evidence required	Decision relevance
Is mapping from score to position fixed and understood?	Ranking, calibration, blending, optimisation and constraint logic	Prevents construction from becoming an unrecorded source of value
Are concentration and tail behaviour acceptable?	Name, group, theme, factor and confidence concentration	Limits loss amplification and hidden bets
Is turnover measured on positions and orders?	Average, tail, burst, retraining and threshold turnover	Captures actual trading rather than score stability alone
Are cost assumptions dependent on market state?	Spread, impact, fees, delay, opportunity cost, financing and borrow where relevant	Tests implementation survival on realistic trades
Is reconciliation of realised costs possible?	Order, fill and benchmark records linked to model version	Enables calibration and attribution after authorisation
Is liquidity evaluated beyond average volume?	Participation, days to trade, days to liquidate and stressed depth	Determines tradability and exit risk
Is capacity presented conditionally?	Net contribution across capital, participation	Prevents capacity from being treated as a

Control question	Evidence required	Decision relevance
	and turnover scenarios	permanent number
Are interactions with existing trades measured?	Internal crossing, conflict, reinforcement and analysis of common trades	Establishes marginal portfolio cost and crowding
Is a construction with lower turnover considered?	Buffer, decay, range in which no trade is made or slower schedule comparison	Tests whether gross edge can be converted more efficiently

B.8 Drift, retraining and change control

Control question	Evidence required	Decision relevance
Are drift types defined?	Feature, label, relationship, portfolio, vendor, parameter and representation measures	Directs diagnosis to the correct source
Is the retraining rule explicit?	Fixed or rolling design, frequency, window, seed and acceptance tests	Prevents discretionary refitting after losses
Are pre-approved changes bounded?	Maintenance actions permitted without a full review	Allows efficient operation without hidden model change
Is materiality assessed by outcome?	Rank, tail, exposure, turnover, cost and contribution comparison	Detects major capital change from minor technical edits
Is a challenger available where useful?	Simple or alternative model with defined comparison rule	Provides diagnostic contrast and fallback information
Are versions complete and immutable?	Code, data rules, features, parameters, artefacts, construction and monitoring	Supports reconstruction, rollback and accountability
Does a material change reset the clock?	Written transition rule for revalidation and shadowing	Prevents successors from inheriting authority informally

B.9 Shadow deployment and operating readiness

Control question	Evidence required	Decision relevance
Is the shadow object frozen?	Version manifest covering the entire decision chain	Ensures that observed behaviour belongs to one object
Does it run on the intended schedule?	Live input receipt, scoring, construction and notional orders	Tests real latency and operating feasibility
Are interventions recorded?	Reruns, corrections, overrides and manual data repairs	Reveals hidden operational dependence
Are failures contained?	Input validation, quarantine, fallback and recovery procedures	Prevents incomplete or stale output from acting
Are proposed trades and costs captured?	Targets linked to the version, orders, liquidity and cost estimates	Tests the implementation case on current conditions
Are dashboard measures actionable?	Owner, expected range, hard limit and response for each key measure	Confirms that monitoring can change the operating state
Is the shadow period preserved after	Reset rule and successor identity	Maintains evidence integrity

Control question	Evidence required	Decision relevance
material change?		
Is the fallback tested?	Successful execution of prior model or simple process	Confirms that suspension is operationally possible

B.10 Permission, ownership and exit

Control question	Evidence required	Decision relevance
Is the authorised function stated narrowly?	Exact portfolio decision and excluded decisions	Prevents authority from expanding by convention
Are capital and action limits external to the model?	Position, factor, gross, liquidity, turnover or action controls	Ensures model conviction cannot override hard limits
Are owners named across the lifecycle?	Research, portfolio, data, technology and independent review responsibilities	Avoids diffuse accountability
Is attribution responsibility assigned?	Review frequency and owner for return, risk and trading contribution	Ensures that continued use is supported by evidence
Are escalation states defined?	Normal, warning, restricted, suspended and killed meanings	Makes operating status unambiguous
Are kill criteria linked to actions?	Trigger, containment, decision owner, fallback and closure condition	Makes withdrawal executable rather than merely stated
Is return controlled?	Required evidence after blocking or killing	Prevents informal reinstatement
Is the permission record complete?	Identity, version, scope, limits, monitoring, change and exit fields	Provides the authoritative statement of current use

Appendix C. Monitoring dashboard for ML signals

The dashboard should be tailored to the authorised function. A ranking signal, sizing model and execution policy require different measures. The structure below is a complete field library, not a requirement to display every measure for every model. Thresholds should be derived from validation evidence, portfolio consequence and operating capability rather than imposed as universal numerical rules.

C.1 Identity and permission header

Every dashboard should begin with the information needed to determine whether the displayed output is the authorised object.

Field	Required content
Model and output identifier	Stable identifier linked to the model inventory and evidence archive

Field	Required content
Current version	Code, data rule, feature, parameter, construction and monitoring versions
Permission state	Shadow, authorised function, restricted, blocked or killed
Authorised use	Exact portfolio decision the output may influence
Scope	Portfolio, universe, horizon, schedule, instruments and eligible actions
Limits	Influence, concentration, exposure, turnover, liquidity, participation or action bounds
Effective and review dates	Start of current authority and next scheduled review
Owners	Portfolio, research, data, technology and independent review owners
Fallback	Approved prior version, rule or operating process
Open decisions	Active warning, restriction, investigation or successor review

C.2 Run data and operations panel

Measure	Comparison	Interpretation and action
Input arrival completeness	Required sources versus received sources	Quarantine the run when material inputs are absent
Freshness and validity at decision time	Source timestamps versus approved threshold	Reject stale inputs where the use depends on timeliness
Universe and coverage	Current eligible names and feature coverage versus expected range	Investigate mapping, vendor or mandate changes
Missingness by source and group	Current pattern versus validation and recent history	Restrict affected features or groups if missingness changes meaning
Revision and restatement activity	Current corrections versus normal process	Assess whether prior scores or training data are affected
Identifier reconciliation	Corporate actions, mappings and eligibility checks	Block outputs that cannot be assigned reliably
Pipeline completion	Approved stages versus completed stages	Use the fallback if the full path did not complete
Latency	Completion time versus decision deadline	Prevent late output from entering a process simulated at an earlier time
Manual intervention	Reruns, overrides, repairs and operator notes	Escalate repeated intervention as an operating weakness
Fallback readiness	Last successful fallback test and current availability	Suspend authority if no safe fallback can be executed

Data and operating failures should generally be easier to contain than statistical deterioration. Where integrity is uncertain, the default should be to stop the affected output rather than guess a live treatment that was not approved.

C.3 Output behaviour panel

Measure	Comparison	Interpretation and action
Score location and dispersion	Current distribution versus validation and recent history	Investigate collapse, shift, saturation or uncontrolled widening
Rank stability	Adjacent runs and history appropriate to the horizon	Detect stale data, excess churn or retraining shock
Tail size and membership	Actionable tails versus expected breadth and overlap	Restrict influence if tail behaviour becomes concentrated or unstable
Missing or invalid outputs	Count, location and reason	Apply the approved treatment; do not improvise
Calibration or confidence	Predicted versus observed outcomes where magnitude matters	Remove sizing or probability authority if calibration no longer supports it
Baseline disagreement	Candidate versus linear or simple policy control	Identify where incremental action is concentrated
Divergence across versions	Current versus proposed or prior version	Require shadowing when portfolio behaviour changes materially
Policy action distribution	Frequency, magnitude and state of actions	Block actions outside the reviewed envelope through external limits

C.4 Evidence and predictive behaviour panel

Measure	Comparison	Interpretation and action
Statistic relevant to the use	Current period versus validation distribution	Review in light of dependence and sample size, not as an isolated pass/fail number
Contribution by date or period	Broad distribution versus concentration in a few intervals	Escalate if value depends on a shrinking set of observations
Contribution by universe segment	Region, sector, size, liquidity and other relevant groups	Narrow scope if errors or value are concentrated unexpectedly
Behaviour conditional on state	Market, volatility, liquidity and crowding states	Reclassify if the model has become a state-timing exposure
Error concentration	Security, group, event and tail clusters	Restrict affected scope and investigate data or construction causes
Benchmark comparison	Authorised model versus simple control	Challenge complexity if incremental value disappears persistently
Challenger comparison	Incumbent versus defined challenger	Use for diagnosis; do not switch based only on recent relative performance

For a risk filter, the relevant evidence may be avoided loss, cost of false positives and missed adverse states rather than an information coefficient. For an execution policy, the relevant evidence may be implementation shortfall, fill quality and action stability. The panel should follow the authorised function.

C.5 Exposure and portfolio panel

Measure	Comparison	Interpretation and action
Beta and aggregate exposure	Score, target and realised portfolio versus limits	Cap, neutralise or reclassify unapproved market exposure
Sector and region	Active weights and risk contribution versus approved range	Restrict concentration outside the permission record
Size, value, momentum and quality	Loadings and realised contribution versus validation	Determine whether intended selection has become style timing
Volatility and low volatility	Average and tail relation versus expected behaviour	Examine defensive or procyclical migration
Liquidity	Holdings, target trades and exit profile versus limits	Reduce authority when tradability deteriorates
Reversal and dependence over short horizons	Current relation and trade pattern versus history	Distinguish return information from contrarian exposure that is sensitive to execution
Crowding and common trade	Position overlap, flow proxies and common exits	Reduce capacity or escalate under concentrated exit risk
Name and theme concentration	Current targets and realised positions versus external limits	Apply hard controls regardless of model confidence
Gross and net contribution	Effect of sizing or overlays on portfolio exposure	Revert if model influence exceeds the authorised corridor
Marginal portfolio contribution	Combined book with and without the model	Assess continued usefulness rather than standalone activity

C.6 Trading, cost and capacity panel

Measure	Comparison	Interpretation and action
Target turnover	Average, tail and cause versus approved budget	Suppress, defer or review trades outside the intended pattern
Realised turnover	Completed trading versus target positions	Investigate construction, crossing or execution differences
Spread and fees	Predicted versus realised	Recalibrate or restrict if errors persist by group or state
Market impact	Predicted versus realised by order size and liquidity	Reduce participation, urgency or capital when impact rises
Delay and opportunity cost	Intended versus completed position path	Review whether signal decay exceeds implementation capability
Borrow and financing	Availability, cost and recall versus assumptions	Block trades whose economics no longer survive
Participation	Current and proposed order use versus approved bounds	Enforce limits outside the forecasting model
Days to trade and liquidate	Portfolio and tail positions versus approved profile	Reduce scale or exclude positions with unacceptable exit time

Measure	Comparison	Interpretation and action
Capacity use	Current capital on the approved scenario curve	Restrict additional capital when marginal value decays
Overlap with common orders	Candidate versus other signals and portfolios	Coordinate, cross or suppress trades that create internal crowding
Implementation shortfall	Realised result versus decision benchmark	Attribute model, construction and execution effects separately

C.7 Attribution panel

Layer	Required decomposition	Decision use
Raw model	Forecast or policy contribution before controls	Determines whether the predictive object remains active
Known exposures	Beta, group, style, liquidity and crowding contribution	Identifies hidden risk and possible reclassification
Portfolio construction	Neutralisation, blending, optimisation, constraints and sizing	Shows whether construction preserves or redirects value
Trading	Spread, impact, delay, opportunity cost, financing and borrow	Measures implementation survival
Interaction	Contribution gained or lost through the surrounding portfolio	Establishes marginal value to the book
Residual	Unexplained contribution after defined layers	Triggers investigation when persistent or concentrated

Attribution should use the version that actually generated each decision. Reconstructing a historical period with the latest model can answer a research question but cannot explain the live portfolio outcome.

C.8 Drift, retraining and version panel

Drift type	Observable measures	Governance response
Feature	Distribution, missingness, coverage, support and source changes	Diagnose cause and downstream effect before retraining
Label	Frequency, scale, horizon behaviour and decision alignment	Revalidate the target and the linked action
Relationship	Conditional forecast error and contribution	Assess decay, reversal or changing state dependence
Portfolio	Exposure, concentration, turnover and cost migration	Restrict or reclassify even if forecast statistics remain stable
Vendor	Schema, mapping, revision and historical restatement	Quarantine affected versions and reconstruct impact

Drift type	Observable measures	Governance response
Parameter	Coefficient, tree, representation or policy change	Compare portfolio behaviour across versions
Representation	Embedding, activation or downstream sensitivity change	Freeze or shadow a successor when economic meaning shifts
Rank and action	Tail overlap, turnover and action divergence	Treat material behavioural change as a new signal
Monitoring	Threshold, measure or change in escalation rule	Version and approve the control environment itself

The panel should distinguish scheduled retraining from material change. A scheduled retrain can remain within authority only when the process and resulting behaviour remain inside the approved range. Otherwise, the successor returns to validation or shadowing.

C.9 Alert and decision log

Every material alert should create a structured record containing the timestamp, model version, permission state, triggering measure, affected data or positions, immediate containment, decision owner, investigation, final decision, effective time, fallback used and conditions for closure.

Warnings should be reviewed collectively where they may share a cause. Repeated small coverage alerts, for example, can indicate a vendor or universe shift even when each incident is individually below a hard limit. An alert should not be closed merely because the next run returned to range if the underlying cause remains unknown.

C.10 Operating status and permitted action

Status	Meaning	Permitted portfolio action
Normal	Evidence and operations remain within the authorised framework	Continue within limits
Warning	A measure requires review but immediate containment is not mandated	Continue only where the permission record permits; owner reviews
Restricted	Authority has been narrowed while evidence is assessed	Operate only within the temporary residual function and limits
Suspended	The output may not affect the portfolio pending decision	Use the approved fallback
Blocked	The output has no authority, though research may continue	Prevent all portfolio influence
Killed	The version is retired	Archive and prevent reactivation; treat any successor as new

The dashboard completes the permission framework. Research establishes a case. Validation tests the case. Shadowing observes the frozen object. Authorisation defines bounded influence. Monitoring determines whether the conditions remain true. Kill rules ensure that influence can be withdrawn when they do not.