All Actuaries Summit 2026
25 – 27 May 2026, Melbourne

Actuaries Institute.

Reputation Risk as a Quantitative Actuarial Practice

Prepared by Colin Priest

Presented to the Actuaries Institute
2026 All-Actuaries Summit
25-27 May 2026

Actuaries Institute
ABN 69 000 423 656
Level 34, Australia Square, 264 George Street, Sydney NSW 2000
P +61 (0) 2 9239 6100 | actuaries.asn.au

Reputation Risk as a Quantitative Actuarial Practice

Abstract

Reputation risk has historically resisted quantitative actuarial analysis because relevant data are sparse, stakeholder responses are strategic, and causal pathways to financial loss are indirect. This paper demonstrates how adversarial risk analysis, generative AI digital twins, and Bayesian estimation can make it tractable, using the Qantas 2023 AGM governance crisis as a case study. The framework correctly identifies the modal outcome at four of five decision nodes and finds that commissioning an independent governance review dominates all alternative Board strategies across all plausible parameter values. The aim is not to claim perfect prediction of unprecedented events, but to provide a replicable actuarial workflow for analysing reputation-driven strategic risk in data-poor settings. All code and tools will be released as open source.

Contents

Practitioner Guide to This Paper

If you are a practising actuary, risk manager, or governance adviser, this guide explains what this paper does, what it produces, and how to navigate it efficiently. Readers primarily interested in the technical methodology should proceed directly to Section 1.

What problem does this paper solve?

Reputation risk is real, it is financially material, and it is already insurable, yet it has resisted actuarial analysis because the data are sparse, the causal pathways are indirect, and the key actors behave strategically. When a governance crisis hits, a board cannot run a regression. It needs to know what its opponents are likely to do, what its own options are worth, and how much it should pay to find out more before acting. This paper demonstrates that adversarial risk analysis, combined with generative AI and Bayesian estimation, can answer those questions in a principled, replicable, and auditable way.

What is the methodology, in plain terms?

The framework has five components, each of which does a specific job.

Stakeholder identification and characterisation. The first step is identifying who has genuine decision power and mapping their objectives, constraints, and likely behaviour from publicly available evidence: governance documents, regulatory filings, historical voting records, and media commentary. This is actuarial data collection applied to non-traditional sources.

Generative AI digital twins. Because key stakeholders are inaccessible and historical precedents are sparse, the framework uses large language models to simulate stakeholder behaviour across many counterfactual scenarios. These digital twins are not treated as ground truth; they are a scalable, auditable way to generate structured preference data that would otherwise be impossible to obtain. Think of them as a computational wind tunnel for governance decisions.

Bayesian estimation of stakeholder preferences. The digital twin outputs are not used directly. They are fed into formal statistical models, ordinal probit for board preferences, random utility models for shareholder advisers, logit-normal models for vote outcomes, that are anchored by historical data and produce full posterior distributions over stakeholder behaviour. The result is not a point estimate but a calibrated probability distribution over what each actor will do.

A sequential game tree. The stakeholder models are connected in a decision tree that mirrors the actual sequence of governance decisions, who moves first, who responds, and what the possible outcomes are at each node. Each path through the tree carries a probability weight, and each terminal node carries an expected utility. The tree is solved by backward induction, the same logic actuaries already use in multi-state insurance models.

Decision outputs. The framework produces expected utilities by action, counterfactual comparisons across alternative strategies, sensitivity analysis over uncertain parameters, and value of information diagnostics. These are the outputs a board or its adviser needs to make a defensible, documented decision under uncertainty.

What did the Qantas case study show?

The framework was applied to the Qantas 2023 AGM governance crisis, in which the Board faced decisions about CEO tenure, governance review, and remuneration strategy under intense public, regulatory, and shareholder pressure. The main findings were:

The framework correctly identified the modal outcome at four of five decision nodes, including the two most consequential, CEO departure and the Board's decision to commission a governance review, without any post-hoc tuning to the observed outcomes.

Commissioning a governance review was the dominant strategy: it produced the highest expected utility whether the CEO resigned or stayed, and regardless of how the shareholder vote resolved. The value of waiting for better information before acting was quantified and found to be negligible.

The cost of inaction was large and measurable. Both available actions produced negative expected utility, but inaction was substantially worse. The decomposition showed that the Board's real exposure lay in the near-certain governance reckoning from review findings, not in vote management or CEO transition optics.

The recommendation was robust. It survived every single-parameter $\pm 50\%$ perturbation in the sensitivity analysis, meaning the Board did not need precise self-knowledge about its own utility weights to act correctly.

What does this mean for actuarial practice?

Reputation risk is not too soft for actuarial analysis. It is a quantifiable, insurable, governance-relevant risk that fits squarely within the profession's existing competency framework. The barriers have been data poverty and methodological immaturity, and generative AI addresses both. The framework demonstrated here is replicable: Section 6.7 provides a step-by-step blueprint, and the open-source code released with this paper means practitioners can adapt it to their own settings without rebuilding from scratch.

The broader lesson is that some reputation crises are action-dominant: the right response is clear across essentially all plausible downstream states, and the actuarial framework's contribution is to demonstrate that robustness rather than to pretend the uncertainty does not exist. Other crises will be information-dominant, where waiting for better signals has genuine strategic value. Knowing which type of crisis you are in is itself a valuable output of the framework, and one that narrative analysis alone cannot produce.

How to navigate this paper

The paper is organised in three layers. Readers who want the full methodology should read Sections 1 through 6 in sequence before turning to the results. Readers primarily interested in the case study and its outputs can read Sections 1, 2, 3.4, and then jump to Sections 7 through 9. Readers who want to replicate the framework in a new setting should focus on Sections 5, 6.7, and 6.8, which together constitute the replication blueprint and modelling architecture summary.

1 What is an Actuary?

1.1 What Actuaries Do, and Why This Matters Here

Across professional definitions (Actuaries Institute, n.d.; International Actuarial Association, 2018), actuaries are consistently characterised as professionals who quantify uncertain future outcomes, evaluate risk trade-offs, and advise organisations under uncertainty. That framing matters here because reputation risk has historically been treated as "too qualitative" for actuarial analysis, even though its consequences are strategic, financial, and governance-relevant.

1.2 Why Reputation Risk is an Actuarial Problem

Reputation risk is typically defined as the risk that actions, events, or associations erode stakeholder trust and thereby destroy organisational value through lost customers, higher funding costs, or regulatory sanctions. Conceptually, it fits within enterprise risk management as a forward-looking, intangible driver that amplifies other risks rather than a stand-alone silo. From this perspective, reputation risk is not "too soft" for actuarial work; it is a correlated, non-traditional risk factor whose effects can be quantified using the same toolkit applied to credit, market, and operational risks.

Reputation risk aligns closely with the established core competencies of the actuarial profession (Actuaries Institute, n.d.) and becomes tractable once suitable data and modelling tools are available. Actuaries are already trained to integrate financial theory, statistics, and data science to evaluate risk and opportunity across insurance, pensions, and broader corporate settings. Professional competency frameworks emphasise precisely the skills needed to make reputation risk analysable: advanced data analytics, scenario analysis, enterprise risk management, and communication with boards and regulators. Viewed in this light, "reputation risk as an actuarial problem" is mainly a question of data and methodology maturity, not of scope creep for the profession.

1.2.1 Core Actuarial Skills Applied to Reputation

Risk management. Actuaries are trained to identify, measure, and mitigate risks to organisational value using structured frameworks such as enterprise risk management and capital allocation. Reputation risk can be decomposed into exposures, perils, and hazards (such as customer-facing channels, social-media firestorms, and governance failures) and assessed in terms of likelihood and severity in exactly the same way as traditional risk categories.

Data analysis. Reputation has historically been "data poor" because relevant signals sit in noisy, sparse, and heterogeneous sources such as media coverage, complaints, analyst reports, and internal whistleblowing channels. Actuaries already work with imperfect insurance and financial data, combining traditional structured data with text, behavioural, and external indicators to extract signals that are decision-relevant. The same techniques (feature engineering, natural language processing, and anomaly detection) can be used to turn reputational narratives into quantitative covariates for loss and survival models.

Uncertainty. Actuaries specialise in quantifying uncertainty, including parameter risk, model risk, and process risk, under incomplete information. Reputation risk exemplifies

Knightian uncertainty: stakeholders' future perceptions are not directly observable and can shift abruptly in response to rare events. Actuarial methods such as Bayesian inference, stress testing, and model-risk assessment provide a disciplined way to articulate what is not known and to embed conservative margins or capital buffers.

Forecasting. Reputation-driven outcomes (policy lapses, deposit run-offs, customer churn, or regulatory interventions) are inherently forward-looking and scenario-dependent. Actuaries already build stochastic projections of cash flows and capital under alternative macroeconomic, demographic, and behavioural scenarios. Extending these to include reputational states or "trust indices" simply adds another layer to the existing projection machinery rather than a fundamentally new modelling paradigm.

Communication. Professional standards emphasise that actuaries must communicate complex analyses clearly to boards, executives, and regulators, including limitations and uncertainties. Reputation risk is board-relevant precisely because it cuts across strategy, conduct, and capital, but decision-makers often lack quantitative framing. Actuaries can translate technical findings, such as the distribution of potential earnings impacts from a conduct scandal, into actionable advice on risk appetite, contingency plans, and investment in controls.

Stochastic optimisation. Many reputational decisions involve trading off short-term gains against long-term trust under uncertainty, for example when setting product pricing, claims-handling policies, or content-moderation thresholds. Actuaries are familiar with stochastic optimisation and portfolio techniques that seek strategies performing well across a distribution of plausible futures, rather than optimising a single point estimate. This is exactly the kind of framework needed to evaluate whether aggressive cost-cutting that increases complaint rates is value-destructive once reputational feedback loops are recognised.

Governance and ethics. The actuarial profession's codes of conduct stress fiduciary duties to the public, regulatory compliance, and the obligation to uphold the reputation of the profession itself. Reputation risk sits at the intersection of governance failures, misaligned incentives, and disclosure choices, all of which are areas where actuaries already interact with boards, audit committees, and supervisors. Their ethical training, including managing conflicts of interest and considering stakeholder impacts, is directly relevant when advising on reputationally sensitive decisions such as product design, restructuring, or AI deployment.

Statistics and model-building. Actuarial education is built on probability, statistics, and modelling, applied to fit models, test hypotheses, and quantify estimation error. The same toolkit can test whether media sentiment Granger-causes changes in sales or funding spreads, or estimate the incremental loss severity associated with high-profile conduct events. Importantly, actuarial standards require explicit documentation of assumptions, validation, and limitations, which is critical in an area where spurious correlations and overfitting are genuine hazards.

1.2.2 Data and Modelling Barriers

Historically, the main barrier to treating reputation risk as an actuarial problem has been the absence of reliable, longitudinal data and standardised modelling tools. Reputation has often been relegated to qualitative "heat maps" in risk registers, precisely because converting narratives and perceptions into structured inputs for capital and pricing models was labour-intensive and ad hoc. As a result, reputational losses showed up ex post in earnings and valuations but were rarely modelled ex ante with the rigour applied to other risk types.

The emergence of advanced analytics, including natural language processing and alternative data sources, has begun to close this gap by enabling systematic harvesting of reputational signals from news, social media, and complaints. However, these techniques remained specialist and often sat outside actuarial teams, limiting their integration into core risk and capital frameworks.

1.2.3 How Generative AI Changes Tractability

Generative AI addresses the data constraint in two principal ways. First, it can transform unstructured text (media articles, customer reviews, analyst reports, regulatory notices) into analysable datasets by extracting entities, events, sentiments, and narratives at scale. This supports the construction of time-series indicators of reputation, labelled event datasets, and features for predictive models that link reputational states to financial outcomes. For actuaries, this means that the raw material needed to quantify reputational drivers becomes available with far lower manual effort, making regular monitoring and back-testing feasible.

Second, generative models can simulate plausible human responses to events and management actions under different contexts, such as alternative crisis-communication strategies or policy changes. While current large language models do not provide "ground truth" behavioural forecasts, they can approximate qualitative patterns of stakeholder narratives and information cascades that are useful inputs to scenario design. These simulated narratives can then be mapped, via actuarial models, into distributions of quantitative outcomes, such as churn rates or litigation frequencies, that are compatible with existing risk and capital frameworks.

At the same time, the use of generative AI itself introduces new reputational vulnerabilities (for example, if AI-generated content is misleading, biased, or perceived as manipulative), which reinforces the need for disciplined, ethically informed risk management. Actuaries' background in model-risk governance, documentation, and transparency positions them to scrutinise where AI-generated data and simulations are appropriate and where human judgment or additional safeguards are required.

1.2.4 Role of Adversarial Risk Analysis

Adversarial risk analysis (ARA) offers a decision-analytic framework for situations where key stakeholders, such as regulators, activists, or coordinated online communities, respond strategically to the organisation's actions. In the context of reputation risk, ARA extends familiar actuarial tools by explicitly modelling these actors' objectives and likely responses, providing decision recommendations that account for strategic behaviour. The next section develops this framework in detail.

1.2.5 Reputation Risk as an Insurable Peril

The insurance market's treatment of reputation risk further confirms its actuarial character. A small but growing number of insurers and specialist managing general agents (MGAs) globally underwrite reputation-focused policies, typically framed as "reputational risk," "reputational crisis," or "reputation value" covers rather than a stand-alone guarantee of an intangible asset. At Lloyd's and in the company market, products from carriers such as AIG (ReputationGuard), Allianz Global Corporate & Specialty (Reputation Protect Plus), AXA XL, and Liberty Specialty Markets, with capacity arranged via brokers like Aon and WTW, combine pre-loss monitoring and crisis-management services with insurance that responds to defined "reputational crisis events," often indemnifying crisis-response costs and, in some cases, loss of gross profit following a trigger event that causes measurable reputational harm.

These covers generally sit alongside, or on top of, more traditional lines (D&O, cyber, product liability) and do not insure "reputation" in the abstract. Instead, they use specific triggers such as adverse media linked to a defined peril, regulatory investigations, or ESG-related controversies, and pay for PR and crisis communications, data and sentiment monitoring, and business-interruption-style losses over a limited post-event period.

Some providers (for example, Steel City Re and certain Lloyd's syndicates) go further and offer parametric or capital-markets-style reputation solutions, where payments are linked to observed indicators of reputation impairment (such as equity price movements or third-party sentiment indices), reflecting a view within the market that reputational value is insurable but requires carefully structured, data-driven triggers to avoid pure subjectivity. This parametric approach aligns naturally with actuarial methods: defining frequency and severity distributions for trigger events, calibrating attachment points and limits from historical data, and pricing the cover via Monte Carlo simulation of loss scenarios. The existence of these products demonstrates that the market has already concluded reputation risk is quantifiable: the question is no longer *whether* it can be measured, but *how well*.

The remaining question is not whether reputation risk can be analysed actuarially, but how to do so in a setting where stakeholder reactions are strategic, data are sparse, and organisational decisions unfold sequentially. The Qantas AGM crisis provides a useful test case because the audience of this paper already knows the broad outcome, allowing the plausibility of the modelling choices to be checked intuitively.

1.3 Roadmap

The paper is organised in three layers. Section 1 explains why reputation risk belongs within actuarial practice. Section 2 introduces the adversarial risk analysis framework used to model strategic stakeholder responses. Sections 3–7 then work through a full Qantas case study, moving from institutional context and stakeholder characterisation to GenAI-assisted elicitation, stochastic modelling, and decision outputs. Readers primarily interested in replication may move directly to Sections 5–6 after reading Section 1 and Section 3.4.

2 What is Adversarial Risk Analysis?

2.1 Definition

Adversarial Risk Analysis (ARA) is a framework for decision-making in situations involving intelligent opponents or strategic stakeholders. As defined in the foundational textbook by Rios Insua, Banks, and Rios (2009):

> Adversarial Risk Analysis provides a unified Bayesian framework for risk analysis in the presence of opponents and uncertain environments. It treats the decisions of other agents as random variables, modelled through the analyst's beliefs about their goals, resources, and capabilities, and integrates these into a coherent decision-theoretic structure.

2.2 Modelling Adversarial Situations Under Uncertainty

Adversarial Risk Analysis (ARA) provides a Bayesian decision-analytic framework for modelling situations in which multiple stakeholders with potentially conflicting objectives act strategically under mutual uncertainty (Rios Insua, Banks, & Rios, 2009; Banks, Rios, & Rios Insua, 2015). In the reputation-risk context developed in this paper, the stakeholders are a corporate Board, a CEO, an activist shareholder association (ASA), a regulator, and a body of retail shareholders. Each stakeholder possesses private information (about their own risk tolerance, strategic priorities, and red lines) that is unobservable to the others. ARA treats every other agent's forthcoming action as a *random variable*, resolved through the focal decision-maker's subjective beliefs about that agent's objectives, constraints, and reasoning process. This contrasts with classical game theory, which requires the common-knowledge assumption (every player knows every other player's utility function, and knows that they know, *ad infinitum*) and seeks Nash equilibria. In corporate governance crises, common knowledge is precisely what is absent: the Board does not know the ASA's true reservation price for governance reform, the CEO does not know the Board's private threshold for requesting a resignation, and shareholders observe only noisy public signals of all parties' intentions (Rios Insua, Rios, & Banks, 2009).

2.2.1 Formal Structure: Actions, Utilities, and Priors

The formal ARA specification requires three ingredients for each stakeholder k: an **action space** $\mathcal{A}_k$, a **utility function** $u_k(a_k, a_{-k}, \theta)$ defined over the agent's own action a_k, the joint actions of all other agents a_{-k}, and an uncertain state of the world θ; and a **prior distribution** $\pi_k(\theta, a_{-k})$ encoding k's beliefs about the state and about what the other agents will do. Uncertainty enters the model at two distinct levels. *Aleatory uncertainty* captures inherently stochastic outcomes (for example, the random variation in proxy-advisery recommendations or the vote-share realisation at an AGM), and is represented by the state variable θ. *Epistemic uncertainty* captures the analyst's incomplete knowledge of each stakeholder's preferences and beliefs, and is represented by prior distributions over utility-function parameters (risk-aversion coefficients, loss-aversion parameters, discount rates, reputational sensitivity weights) that are updated via Bayesian inference as data accumulate (Rios Insua, Banks, & Rios, 2009). This two-level uncertainty structure is central to the actuarial value of the framework: it forces

explicit quantification of what is known, what is uncertain, and how residual uncertainty propagates into decision recommendations.

2.2.2 Posterior Computation and Predictive Action Distributions

Given a specification of priors and utilities, the analyst solves each stakeholder's decision problem *from that stakeholder's perspective*, integrating over the analyst's uncertainty about the stakeholder's parameters. For stakeholder k, the optimal action under the analyst's beliefs is:

$$a_k^* = \arg\max_{a_k \in \mathcal{A}_k} \mathbb{E}_{\pi_k}\big[u_k(a_k, a_{-k}, \theta)\big],$$

where the expectation is taken over the joint prior $\pi_k(\theta, a_{-k})$. Because π_k itself depends on the analyst's uncertain beliefs about k's utility parameters, the predictive distribution over k's action is obtained by a further integration over the posterior distribution of those parameters. In practice, these nested integrals are intractable in closed form, and are evaluated via Markov chain Monte Carlo (MCMC) methods, specifically Hamiltonian Monte Carlo (HMC) implemented in Stan, which produce correlated draws from the joint posterior and thereby yield Monte Carlo estimates of the predictive action distributions (Ekin, Naveiro, Rios Insua, & Ruggeri, 2021).

2.2.3 Game-Tree Structure and Sequential Resolution

ARA structures multi-agent interactions as a **game tree** in which each node corresponds to a stakeholder's decision point and each edge corresponds to a possible action. At every node, the acting stakeholder's choice is modelled not as a deterministic best-response (as in subgame-perfect equilibrium) but as a *draw from the predictive action distribution* described above. This means that every path through the tree carries a probability weight equal to the product of the predictive probabilities along its edges, and the focal decision-maker's expected utility is computed by averaging over all paths. For multi-step games with several adversaries (precisely the structure of the Qantas governance crisis, where the Board, CEO, ASA, and shareholders make sequential decisions over multiple rounds), Ekin et al. (2021) developed augmented probability simulation methods that make these computations tractable by exploiting common random numbers and variance-reduction techniques.

2.2.4 Operationalisation for the Qantas Case Study

In the Qantas application developed in Sections 4–7, the game tree covers the 2023 AGM cycle and its aftermath, with the following decision sequence: (i) the CEO (D_0^{ceo}) decides whether to resign voluntarily or remain in position; (ii) the Board (D_1) decides whether to take no pre-emptive action, commission an independent governance review, or force CEO transition; (iii) the ASA (A_2) decides whether to recommend voting against the remuneration report; (iv) shareholders vote on the remuneration report under the two-strikes rule (Nature node V); (v) the Board responds post-AGM (D_{rev}) and a governance review, if commissioned, produces findings (Nature node R); and (vi) in the counterfactual branch where the CEO stayed, the CEO faces a further post-review decision (D_4'): stay, resign late, or negotiate an exit, and the Board may act post-review by removing the CEO. Each stakeholder's utility function is parameterised and estimated from observable data (remuneration reports, ASA public

statements, proxy-adviser recommendations, AGM voting records) using the Bayesian methods described in Section 6. The resulting model produces a joint probability distribution over all governance outcomes (CEO departure, board composition changes, remuneration policy reforms, regulatory intervention), conditional on each stakeholder's strategic choices, enabling the actuarial analyst to quantify reputation risk in probabilistic terms.

2.2.5 Actuarial Coherence

The ARA framework satisfies the axiomatic foundations of Bayesian decision theory (Rios Insua, Banks, & Rios, 2009), which guarantees internal coherence: the resulting risk quantification is consistent with the analyst's stated beliefs and preferences, and updates rationally as new information arrives. This property is essential for actuarial applications, where internal consistency of assumptions is a professional and regulatory requirement. Moreover, because the framework produces full predictive distributions rather than point estimates, it naturally supports the calculation of standard actuarial risk measures (Value-at-Risk, Tail Value-at-Risk, and scenario-conditional expectations) over governance and reputation outcomes.

2.3 Successful Real-World Applications

ARA has been applied successfully in a range of high-stakes domains. Banks et al. (2015) provide a comprehensive survey of the field's development and applications; the examples below illustrate the breadth of problems to which the framework has been applied.

- **Counter-terrorism and homeland security.** The original motivation for ARA arose from the limitations of probabilistic risk assessment (PRA) for intelligent adversaries. Merrick and Parnell (2011) compared PRA and ARA methods for counterterrorism resource allocation, showing that ARA produces materially different, and more defensible, defensive postures when attackers can observe and adapt to the defender's strategy. The US Department of Homeland Security adopted ARA-based models to allocate screening resources across airports, border crossings, and critical infrastructure, where the key insight is that adversaries shift their targeting in response to observable defensive investments. Traditional risk matrices treat attack probabilities as exogenous; ARA treats them as endogenous to the defender's choices, producing resource allocations that are robust to adversarial adaptation.

- **Cybersecurity.** Rios Insua, Couce-Vieira, Rubio, Pieters, and Grossklags (2021) developed a comprehensive ARA framework for cybersecurity risk management, modelling the interaction between network defenders and sophisticated attackers with unknown capabilities, resources, and objectives. Their framework decomposes cyber risk into sequential attack–defend–attack games where the defender must anticipate the attacker's reconnaissance, exploit selection, and lateral movement decisions without knowing the attacker's skill level or motivation (financial, espionage, disruption). Couce-Vieira, Insua, and Kosgodagan (2020) extended this to forecasting the economic impacts of cyberattacks, using ARA to model attacker decision-making under uncertainty about the defender's detection capabilities. A central advantage over game-theoretic approaches is that ARA does not require the

common-knowledge assumption: the defender need not assume the attacker knows the defender's utility function, and vice versa.

- **Auction design and competitive bidding.** Rios Insua, Rios, and Banks (2015) applied ARA to first-price sealed-bid auctions, where each bidder has private valuations and must form beliefs about competitors' bidding strategies without assuming Nash equilibrium. The ARA approach models each competitor's bid as a random variable drawn from the focal bidder's subjective belief distribution, calibrated from historical bidding data and domain knowledge. This produces bidding strategies that are robust to misspecification of competitor behaviour, a critical advantage in procurement and spectrum auctions where the common-knowledge assumptions of classical auction theory are routinely violated.

- **Sequential adversarial games.** Ekin et al. (2021) developed augmented probability simulation methods for multi-stage sequential games, where the computational cost of nested belief hierarchies ("I think that you think that I think...") grows exponentially with the number of stages. Their methods enable tractable ARA solutions for games with three or more sequential decision points and multiple adversaries, precisely the structure encountered in the Qantas case study, where the Board, CEO, ASA, and shareholders make decisions in sequence over multiple rounds.

- **Adversarial classification and machine learning.** Naveiro, Redondo, Rios Insua, and Ruggeri (2019) extended ARA to adversarial classification problems, where an attacker manipulates input data to cause a classifier to produce incorrect outputs (e.g. spam filtering, fraud detection, intrusion detection). The ARA framework models the attacker's manipulation strategy as a function of their beliefs about the classifier's decision boundary, producing classifiers that are robust to strategic data manipulation, a problem that classical machine learning treats as a fixed distribution shift rather than an adversarial interaction.

- **Maritime security and piracy.** Wang and Rios Insua (2019) applied ARA to maritime piracy in the Gulf of Aden, modelling the interaction between naval patrols and pirate groups who adaptively select targets based on observed patrol patterns. The ARA model captured the pirates' resource constraints, risk tolerance, and information-gathering capabilities, producing patrol strategies that reduced successful hijackings by accounting for the adversary's strategic response to defensive deployments.

In each case, ARA's strength lies in its ability to handle **deep uncertainty** (Walker, Lempert, & Kwakkel, 2013) (situations where the analyst does not know the opponent's utility function, information set, or decision rule) and still produce actionable recommendations grounded in probability and decision theory (Rios Insua, Rios, & Banks, 2009). The framework's Bayesian foundations mean that prior beliefs about opponents are updated as new information arrives, and the resulting recommendations degrade gracefully as uncertainty increases (wider posterior predictive intervals, more hedged strategies) rather than failing catastrophically as Nash equilibrium solutions do when common-knowledge assumptions are violated.

3 Case Study: Qantas

The modelling choices in Sections 5 and 6 are grounded in evidence, not assumption. The utility weights, prior distributions, and stakeholder behavioural parameters used throughout the ARA framework are calibrated against a decade of observable Qantas conduct, public responses, and governance decisions. Section 3 documents that evidence base.

The chronological narrative that follows serves three modelling purposes. First, it establishes the cumulative reputational context that shaped every stakeholder's priors entering the November 2023 AGM: the Board's awareness of its own governance failures, the ASA's assessment of management accountability, and shareholders' accumulated distrust. Second, it provides the institutional backdrop for the two-strikes rule and the specific governance mechanisms through which stakeholder power was exercised. Third, it introduces the reader to the Qantas case at the level of factual detail needed to evaluate the plausibility of the modelling choices made in Sections 5 and 6, including, in particular, the CEO's prior probability of resignation and the Board's utility function parameters.

Readers primarily interested in the modelling pipeline may move directly to Section 5 after reading Section 3.4, which defines the actuarial framing adopted for the remainder of the paper. Those who engage with the full chronology will find that the modelling choices in later sections follow directly from the institutional record documented here.

3.1 A Decade of Reputation Damage

Qantas Airways, Australia's flag carrier, experienced a sustained sequence of reputation-damaging events between 2010 and 2023 that collectively eroded public trust, employee morale, and political goodwill. The chronology below documents each major episode, the company's public response, and the reputational consequences.

3.1.1 2010: QF32 Engine Failure and A380 Grounding

On 4 November 2010, Qantas Flight 32, an Airbus A380 from Singapore to Sydney, suffered an uncontained failure of a Rolls-Royce Trent 900 engine shortly after take-off, causing serious structural and systems damage. The aircraft landed safely and no one was injured. The Australian Transport Safety Bureau (ATSB) traced the cause to a manufacturing defect in an oil feed stub pipe that led to fatigue cracking, an oil fire, turbine disc fracture, and high-energy debris penetrating the wing and fuselage.

CEO Alan Joyce announced the immediate grounding of the entire A380 fleet, stating that aircraft would not return to service until Qantas was "absolutely confident" about safety, framing the decision as putting safety ahead of commercial considerations. Media coverage initially focused on the near-disaster and praised the crew, while raising questions about Rolls-Royce quality control. Because the defect was attributed to Rolls-Royce manufacturing rather than Qantas maintenance or operations, there was no significant regulatory penalty, and the reputational impact was short-lived and largely mitigated by the safety-first narrative.

3.1.2 2011: Industrial Disputes and Unprecedented Fleet Grounding

Throughout 2011, Qantas became embroiled in bitter disputes with three unions (engineers, baggage and catering staff, and long-haul pilots) over pay, conditions, and

concerns about offshoring and the creation of new Asian-based operations. After months of rolling, legally protected industrial action that disrupted schedules and was estimated to cost Qantas around A$68 million, Joyce announced on 29 October 2011 the immediate lockout of employees covered by the three unions and grounded the entire mainline fleet worldwide.

Joyce argued that the grounding was necessary to protect the airline's long-term viability and to bring the dispute to a head, describing it as a decisive step in the face of "unions' industrial campaign." Unions and politicians characterised the move as "militant management" and a pre-planned ambush of workers and customers. Media reaction was highly critical, highlighting 68,000 stranded passengers, disruption at 22 airports, and damage to Australia's economic interests, casting the grounding as a reputational self-inflicted wound rather than a last resort. The federal government applied to Fair Work Australia, which ordered all industrial action (including Qantas's lockout) to cease. While this enabled operations to resume, the episode entrenched a narrative of combative management and significantly damaged Qantas's brand and customer goodwill.

3.1.3 2013–2014: Job Cuts, Financial Losses, and Brand Erosion

From 2013 into 2014, Qantas announced large losses and plans to cut thousands of jobs as part of a major cost-reduction and restructuring program, feeding a narrative of decline and damaging its image as a stable national carrier. Joyce defended the restructuring as essential to respond to intense international and domestic competition, describing the cuts as painful but necessary to secure Qantas's future. Media and political reaction highlighted staff anger, union resistance, and concerns that service quality was deteriorating. Analysts noted that repeated restructuring announcements and workforce reductions undermined employee morale and public trust, and that the share price reflected a combination of financial stress and questions about management strategy. Although not tied to a specific legal finding, this period is often cited in retrospective analyses as the start of a longer reputational slide, with job cuts and labour conflict recurrent in later controversies.

3.1.4 2020: COVID-19 Response, Mass Stand-Downs, and Outsourcing

In March 2020, as COVID-19 travel restrictions took hold, Qantas suspended all international flights, cut around 60 per cent of domestic capacity, stood down approximately two-thirds of its workforce, and later announced plans to permanently cut 6,000 jobs as part of a pandemic survival and recapitalisation strategy. While Joyce and Qantas framed these measures as unavoidable in the face of unprecedented border closures and revenue collapse, unions and commentators argued that staff were being sacrificed while the airline received substantial government support and management preserved its strategic agenda.

In November 2020, Qantas decided to outsource ground handling at 10 airports, making around 1,700 workers redundant. Management claimed "sound commercial reasons" and cost savings of approximately A$100 million annually, but unions alleged the move was designed to undercut collective bargaining and industrial power. Media coverage and political debate increasingly portrayed Qantas as prioritising cost-cutting and shareholder interests over loyalty to long-serving staff, laying the groundwork for later reputational damage when courts examined the outsourcing decision.

3.1.5 2020–2023: Unlawful Outsourcing and High Court Loss

The Transport Workers' Union challenged the 2020 outsourcing in the Federal Court, which in 2021–2022 found that Qantas had breached the Fair Work Act by outsourcing roles to prevent employees from exercising their rights to engage in future protected industrial action. Qantas appealed unsuccessfully to the Full Federal Court and then to the High Court, which in September 2023 unanimously upheld that the airline had illegally laid off approximately 1,700 ground staff at 10 airports.

In public statements after the High Court decision, Qantas said it "deeply regret[ted] the personal impact" of the outsourcing, reiterated that there had been "sound commercial reasons," and "sincerely apologise[d]" to those affected, while committing to compensation and penalties to be determined by the Federal Court. Media commentary described the judgment as a severe blow to an already damaged corporate reputation, confirming public suspicions that Qantas had treated long-serving workers unfairly and used legal manoeuvres to delay accountability. The decision fuelled calls for board and leadership accountability and was cited in analyses of Qantas's share-price weakness throughout the second half of 2023.

3.1.6 2022–2023: Post-COVID Service Failures and Political Scrutiny

As demand returned in 2022, Qantas attempted to ramp up operations quickly, but customers experienced high cancellation rates, delays, lost baggage, and poor call-centre performance, triggering widespread complaints and negative media coverage. Qantas subsequently acknowledged that it had "let customers down during the post-COVID restart," with management citing labour shortages and operational upheaval, and Joyce issued apologies while emphasising efforts to restore reliability and invest in customer service.

Media and public commentary increasingly portrayed the airline as having allowed standards to slip while executives enjoyed strong remuneration and the company benefited from government support. The airline's reputation metrics deteriorated sharply (Figure 2), and political attention intensified, culminating in a parliamentary inquiry into the airline sector in 2023 in which Qantas's behaviour featured prominently. The share price fell from around A$6.80 in March 2023 to approximately A$4.74 by late October 2023, with sustained negative sentiment attributed to a cluster of reputational issues including service failures, labour disputes, litigation losses, and regulatory actions.

3.1.7 2023: "Ghost Flights" and ACCC Federal Court Action

On 30 August 2023, the Australian Competition and Consumer Commission (ACCC) commenced Federal Court proceedings alleging that Qantas had engaged in false, misleading, or deceptive conduct by advertising and selling tickets on flights it had already cancelled. The ACCC claimed Qantas kept some cancelled flights on sale for an average of more than two weeks and in some cases up to 47 days, and that for more than 10,000 flights scheduled between May and July 2022, the airline delayed notifying existing customers of cancellations for an average of 18 days.

In its public response, Qantas said it "fully accepts it let customers down during the post-COVID restart" and admitted that "mistakes were made," but argued that the ACCC case "ignores the realities of the aviation industry," stressing that airlines cannot guarantee exact flight times and that most affected passengers were offered

near-equivalent alternatives. Media reports labelled the allegation the "ghost flights" scandal and described the ACCC action as another major blow to Qantas's already damaged reputation, reinforcing narratives of customer deception and price gouging. The case created the prospect of substantial civil penalties (public discussion referred to potential fines in the hundreds of millions of dollars) and contributed to the roughly 30 per cent decline in Qantas's share price between March and late October 2023.

3.2 The Annual General Meeting Approaches

With the annual AGM approaching in several weeks, the stakes are high. The AGM provides shareholders with their most direct mechanism of accountability over the Board and executive leadership. Understanding the institutional architecture of that mechanism, the "two-strikes" rule, is essential context for the game tree that follows.

3.2.1 The Two-Strikes Rule

Australian listed companies are subject to the "two-strikes" regime on remuneration votes, contained in Part 2G.2 of the Corporations Act 2001. Each year, shareholders vote on a non-binding resolution to adopt the remuneration report. If 25 per cent or more of votes cast are against the report, the company receives a "first strike." If at the following AGM 25 per cent or more again vote against the remuneration report, this "second strike" automatically requires a separate "spill resolution" at that same AGM, asking shareholders whether to hold a further meeting at which all directors (other than the managing director) must stand for re-election. If the spill resolution passes by a simple majority, the company must convene the spill meeting within 90 days, and board seats can then be vacated and refilled in accordance with the shareholders' votes.

The policy rationale, when the regime was introduced in 2011, was to strengthen board accountability for executive pay and to give shareholders a credible enforcement mechanism short of direct regulation of pay levels. By turning the say-on-pay vote from a purely advisory signal into one with potential consequences for directors' tenure, the regime encourages better alignment between pay, performance, and shareholder interests. In practice, the two-strikes rule also provides a focal point for broader governance concerns: institutional investors and proxy advisers often use the remuneration vote to express dissatisfaction not only with pay, but with strategy, board composition, or responsiveness to prior feedback, increasing the reputational and practical cost to boards of ignoring significant dissent.

3.2.2 Empirical Evidence: Strikes as Threat Rather Than Sanction

Historically, first strikes have been reasonably common but actual board spills rare, suggesting the rule operates mainly as a credible threat rather than a frequently executed sanction. Early analyses found that around 6 per cent of meetings recorded a strike in the regime's first three years; dozens of companies received two strikes, but only a handful were ever forced to hold spill meetings, and no incumbent director of a large listed company lost their seat directly as a result of a spill.

Empirical and practitioner analyses nonetheless attribute tangible effects to the regime. Studies show that CEO pay tends to fall or be restructured after a strike, and boards now devote substantially more disclosure and consultation effort to remuneration design, even though the formal mechanics of spills are infrequently invoked. Even

"minority strikes" (just above the 25 per cent threshold) are associated with subsequent reductions in abnormal CEO pay and reputational penalties for independent directors, such as higher turnover and loss of outside directorships, indicating that boards treat even a relatively narrow strike as a serious signal they cannot ignore.

Reasons commonly cited for "against" votes include weak alignment between incentives and long-term performance, undisclosed or discretionary bonuses, poor handling of misconduct or risk failures, and concerns about board responsiveness or independence. The remuneration report thus becomes a convenient coordination device for broader governance protest, precisely the dynamic at play in the Qantas case.

3.2.3 Post-Strike Remediation

After a first strike, boards typically treat the next year's remuneration report as a remediation exercise and make visible changes across structure, outcomes, and disclosure. Common adjustments include:

- Reducing fixed pay or incentive opportunity for the CEO and key executives.
- Tightening performance hurdles by adding relative TSR or ROE tests, lengthening vesting periods, or removing "soft" metrics.
- Cutting or withholding bonuses where performance or risk outcomes were poor, and removing controversial one-off awards or termination benefits.
- Introducing or strengthening malus and clawback provisions, rebalancing pay more heavily towards long-term equity, and adjusting comparator groups where investors or proxy advisers had criticised them as too generous.

Companies also typically overhaul how they explain and govern remuneration. The Corporations Act requires the next remuneration report after a first strike to explain whether and how shareholder concerns have been taken into account. Boards respond by expanding the discussion of the consultation process, summarising themes from investor feedback, and explicitly linking each design change to those concerns. Many issuers increase direct engagement with investors and proxy advisers in the lead-up to the next AGM, add clearer scenario charts, "single-figure" pay tables and worked examples, and sometimes change committee membership or chair roles to demonstrate refreshed oversight.

3.2.4 Pre-AGM Engagement and Strike Avoidance

Companies try to avoid a first strike by combining **substance** (changes to pay and governance) with **process** (engagement and disclosure). On the substance side, boards stress-test pay outcomes against performance, pre-emptively cut or forgo bonuses if returns or risk outcomes have been weak, and avoid controversial elements such as large sign-ons, retention grants without performance hurdles, or generous termination benefits when there has been a perceived failure. Boards increasingly align variable pay with metrics investors favour (relative TSR, ROE, cash-flow, and non-financial risk measures) and benchmark structures against peer practice and proxy-adviser guidelines.

On the process side, the most effective pre-AGM engagement tactics are early, targeted, and two-way. Boards that start engagement well before the notice of meeting is finalised,

focus on their largest voting shareholders and key proxy advisers, and enter those meetings with concrete data and potential compromises tend to avoid surprises on the remuneration vote. Good practice guidance emphasises mapping who actually votes (including how much of the register is influenced by each proxy adviser), stress-testing proposed pay outcomes against investor and proxy guidelines, and explaining clearly how the structure supports long-term value, risk management, and strategy. Monitoring early proxy returns and nominee instructions allows companies to intensify outreach and clarify contentious points before votes are finalised.

Investors and governance bodies consistently report that direct access to the chair of the board or remuneration committee, detailed but plain-English explanations of metrics and outcomes, and visible willingness to adjust contentious elements are more persuasive than generic presentations. Companies facing heightened risk of a strike, in sectors under conduct or performance pressure, have successfully avoided it by pre-committing in those engagements to lower CEO bonuses, adding harder performance hurdles, or tightening malus and clawback arrangements, and then clearly flagging those changes in the remuneration report as a response to shareholder concerns.

3.2.5 Independent Remuneration Reviews

Companies quite often commission external or "independent" remuneration reviews as part of managing the risk of a first-strike vote, especially where they anticipate controversy or have already seen elevated dissent on earlier resolutions. Boards bring in external remuneration consultants to benchmark pay against peers, stress-test incentive structures against investor and proxy-adviser expectations, and provide an arm's-length assessment that can be cited in engagement meetings and the remuneration report. This is most common in sectors under political or regulatory scrutiny on conduct and culture (e.g. financial services after the Royal Commission), where an external review helps demonstrate that the board has taken concerns seriously and is willing to adjust legacy arrangements.

For investors, these reviews are more persuasive when they have clear scope (e.g. reconsidering metrics, vesting, and termination benefits), the consultant is perceived as independent of management, and the company discloses both key findings and specific changes made in response. External reviews are not a shield if boards ignore their recommendations, but they can be effective in reducing "against" votes when combined with genuine pay changes and early engagement, particularly where the review leads to lower quantum, stronger long-term alignment, or removal of controversial discretion that had driven prior dissent.

This institutional machinery (the two-strikes rule, the engagement cycle, and the independent review mechanism) forms the backdrop against which the Qantas Board must make its strategic decisions as the November 2023 AGM approaches.

3.3 The Australian Shareholders Association

The Australian Shareholders Association (ASA) has been actively engaging with Qantas in the lead-up to the AGM. The ASA has met with Qantas management and board representatives, asking pointed questions about governance, executive accountability, and the company's response to the ghost flights scandal and other reputation-damaging events.

3.4 Your Role

The events described in this case study are real, but the actuarial engagement is a hypothetical framing adopted for illustrative purposes: Qantas did not, in fact, retain an actuary to advise on reputation management during this crisis. For the remainder of the paper, we invite the reader to imagine that **you are an actuary who has been hypothetically retained by the Qantas Board** to advise them on reputation management strategy. Your task is to apply adversarial risk analysis to model the decisions of the key stakeholders, quantify the likely outcomes under different strategic choices, and recommend a course of action that maximises the Board's expected utility while accounting for the uncertain and strategic behaviour of all parties.

Figure 1 summarises the workflow used throughout the rest of the paper, from case evidence to stakeholder modelling to final Board recommendations.

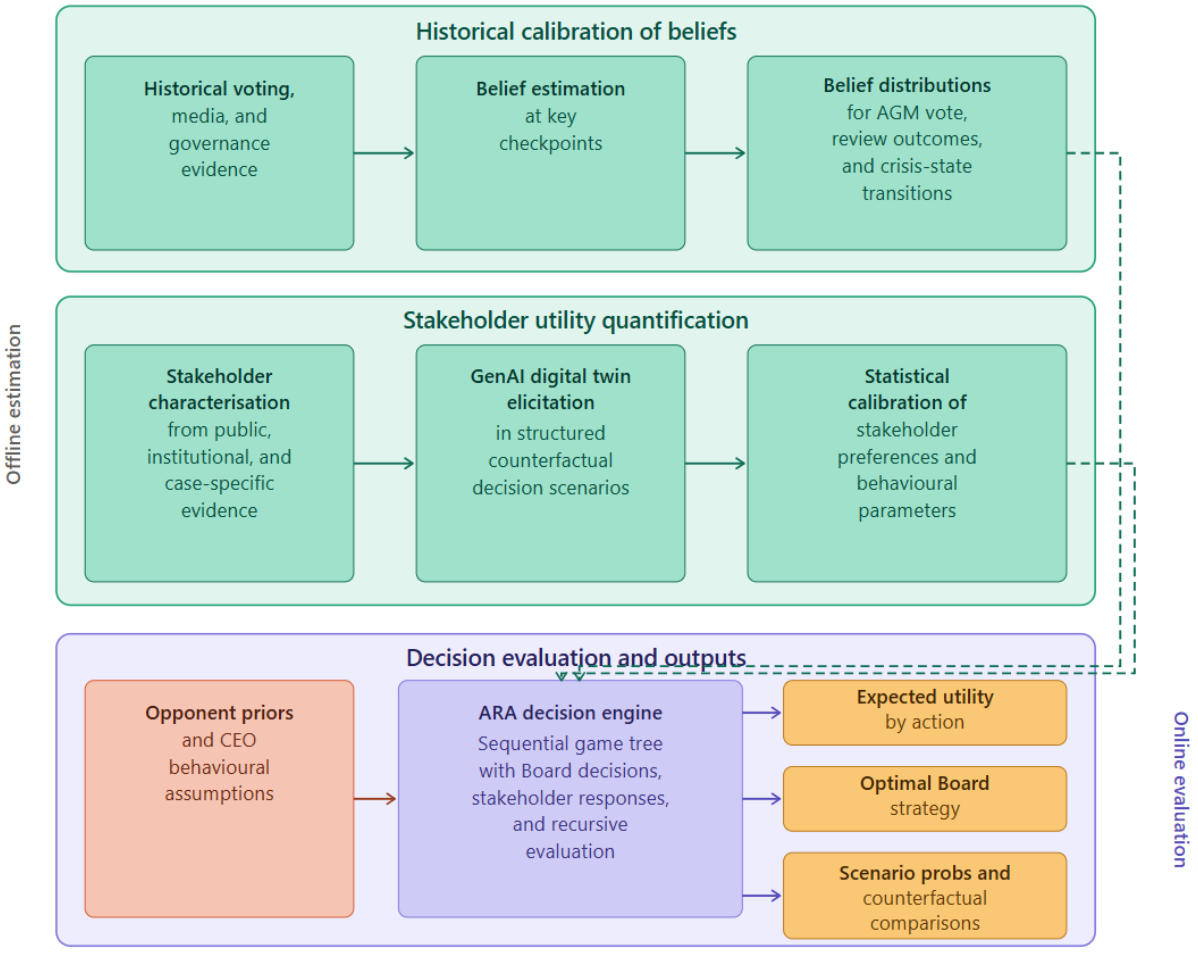

Figure 1: End-to-end workflow for GenAI-assisted adversarial analysis of reputation risk. Historical voting, media, and governance evidence are used to estimate belief distributions at key checkpoints. In parallel, stakeholder evidence is converted into structured GenAI digital twin elicitation tasks, which are then statistically calibrated into preference and behavioural inputs. These belief and utility components feed a sequential adversarial risk analysis (ARA) decision engine, which evaluates alternative Board actions and produces expected utilities, recommended strategies, and scenario-level counterfactual outputs.

4 The Stakeholders

Each stakeholder subsection below follows a consistent four-step template: (i) observed institutional features, (ii) implied objectives, (iii) implied likely actions, and (iv) model entry point; Table 1 consolidates these results across all stakeholders.

4.1 Customers

The Australian domestic airline market is an **oligopoly with fixed supply**. Qantas holds approximately two-thirds of the domestic market share, with Virgin Australia as the only significant competitor. Customers have limited ability to switch: routes, schedules, and loyalty program lock-in constrain their choices. Customer dissatisfaction therefore translates slowly and incompletely into revenue loss. The primary channel of customer influence is through media sentiment and social licence, rather than direct market discipline.

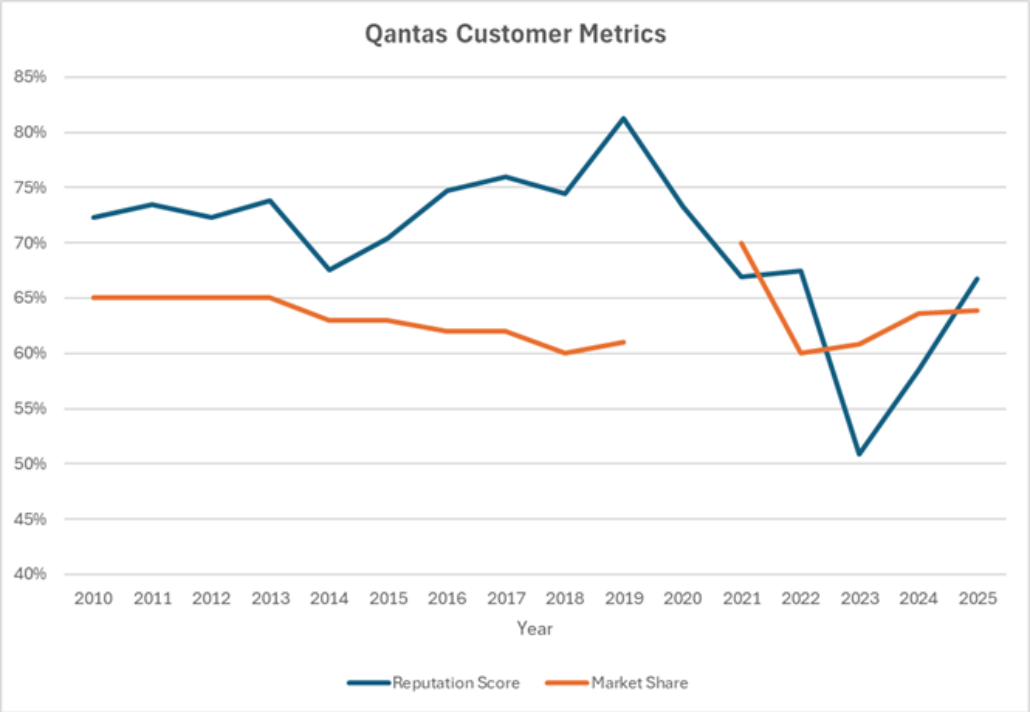

Figure 2: Qantas reputation score and domestic market share, 2010–2025. The reputation score collapsed from 81% (2019) to 51% (2023) following the pandemic-era service failures and ghost flights scandal, while market share remained stable at 60–65%, illustrating the oligopoly insulation that limits the translation of customer dissatisfaction into competitive loss.

Due to the lack of competition, Qantas was able to maintain its market share despite a historical low in its reputation score caused by on-going reputation-damaging events. An analysis of Qantas share price movements over the decade leading up to 2023 suggests that high-profile reputation-damaging events are poor predictors of short-term share price movements.

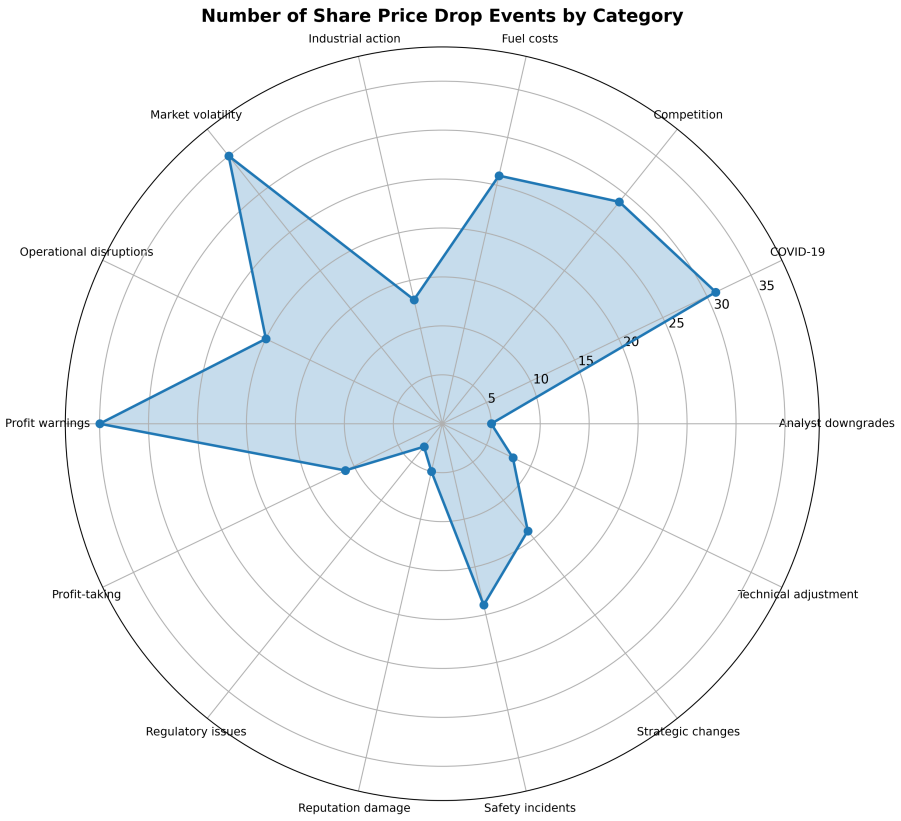

Figure 3: Number of Qantas share price drop events by category, 2013–2023. Profit warnings, market volatility, and COVID-19 dominate the distribution, while reputation damage and regulatory issues account for a small fraction of observed drops, consistent with the weak link between reputational events and short-term share price movements.

Australia's domestic aviation market has long been characterised by a highly concentrated duopoly dominated by the Qantas Group and Virgin Australia, which the ACCC has repeatedly found delivers insufficient competitive pressure to fully meet consumers' needs (Australian Competition and Consumer Commission, 2025). ACCC monitoring reports note that attempts by newer carriers to challenge this structure have not been sustained, reinforcing concerns that limited competition contributes to higher airfares and fewer choices for passengers. In parallel, the federal government and the ACCC have highlighted significant gaps in aviation-specific consumer protections, with many passenger rights currently relying only on general provisions of the Australian Consumer Law. In response, recent policy processes, including the Aviation White Paper and subsequent consultation on aviation consumer protections, propose a dedicated framework built around a Charter of Customer Rights, an Aviation Consumer Ombudsperson, and a new Aviation Consumer Protection Authority (ACPA) to enforce standards and resolve disputes. These initiatives reflect a growing recognition that, in a market with persistent structural concentration and limited effective rivalry, a specialised regulator is needed to protect customers, lift service standards, and provide independent redress when airlines fail to meet their obligations.

For this reason, customers are treated in this paper as an important background reputational force rather than as a strategic decision-maker in the game tree. Their dissatisfaction matters because it shapes media tone, political pressure, and the broader legitimacy environment, but in the November 2023 governance crisis they do not have a

direct lever comparable to the Board, ASA, proxy advisers, or voting shareholders. The radar chart is therefore not a decorative summary; it is part of the modelling justification for excluding customers as a primary strategic actor.

4.2 The CEO: Alan Joyce

Alan Joyce served as Qantas Group Chief Executive Officer from 2008 until his early retirement in September 2023, following a period of intense public, regulatory, and political scrutiny of the airline's conduct and performance. His tenure combined major strategic and financial milestones with a series of decisions that attracted sustained criticism from unions, commentators, and governance analysts, culminating in a formal governance review that identified leadership style and Board–management dynamics as root causes of Qantas's loss of stakeholder trust (Qantas Airways, 2024).

4.2.1 Governance Review Findings

The Qantas Governance Review found that Qantas operated with a "command and control" leadership style centred on "a dominant and trusted CEO," which contributed to "top-down leadership ... insufficient listening and low speak up." The report concluded that there was "too much deference to a long-tenured CEO who had endured and overcome multiple past operational and financial crises," and that this culture led, at times, to a focus on financial performance ahead of non-financial stakeholder risks (other than safety) and to combative external communications that exacerbated issues. As part of its response, the Board later reduced Joyce's remuneration by approximately A$9.26 million, explicitly linking this to findings that executive and Board performance had contributed to reputational damage and customer service failures (Australian Institute of Management, 2024).

4.2.2 Media and Investor Commentary

Media and investor commentary over the past decade repeatedly characterised Joyce as a highly dominant leader who was difficult to challenge and slow to acknowledge error. An analysis of Qantas's "ethical collapse" described the Board under Joyce as "servile," arguing that he shifted from "a competent but arrogant CEO to a self-entitled princeling with a Messiah complex," and that his intolerance of critics and concentration of power amounted to "running a dictatorship." The same piece pointed to episodes such as the 2011 fleet grounding, the lobbying against critical journalism, and the banning of a major financial newspaper from Qantas lounges as "warning[s] to shareholders" about the risks of an ego-driven leadership style. Other commentators noted that Joyce appeared "impervious to criticism and accountability," and that the eventual "pile-on" came only after years in which concerns about culture, service, and workforce relations had been largely discounted (Intelligent Investor, 2023).

4.2.3 Lack of Contrition and Union Criticism

Criticism also focused on the perceived lack of contrition in the face of adverse findings and customer anger. Qantas issued public apologies during the post-COVID operational disruptions, but union leaders argued that the airline was trying to "buy off forgiveness" and called for Joyce's resignation, citing illegal outsourcing, wage suppression, and

customer service failures as symptoms of deeper leadership problems. In 2023, after Qantas lost its High Court appeal over the outsourcing of nearly 1,700 ground staff and the ACCC filed proceedings alleging misleading conduct over "ghost flights," critics framed Joyce's departure as an overdue accountability moment rather than a routine succession. The governance review subsequently echoed aspects of these critiques by emphasising that leadership culture and communication style had underpinned several of the events that damaged Qantas's reputation (ABC News, 2022).

4.2.4 Modelling Implications

For modelling purposes, the relevance of this material is not that media characterisations are taken as literal truth, but that they shape a consistent prior over the CEO's strategic behaviour. The evidence above suggests four modelling implications:

(i) a strong preference for control and resistance to externally imposed accountability;

(ii) a tendency to prioritise financial and narrative management over conciliatory signalling;

(iii) a lower prior probability of voluntary contrition absent severe pressure; and

(iv) a higher likelihood that other stakeholders will anticipate defensiveness and escalate accordingly.

These implications are used to parameterise the CEO digital twin and the opponent priors in Section 6.5 (The Sydney Morning Herald, 2024).

4.3 Shareholders

Shareholders have limited direct power over the Board or the CEO in the ordinary course of business. However, several features of the Qantas shareholder base are relevant:

- Many of the **institutional shareholders are industry superannuation funds** (such as HESTA, AustralianSuper, and Cbus) that have strong ethical and governance reputations to maintain.
- These funds have a **history of shareholder activism**, particularly on ESG (Environmental, Social, and Governance) issues.
- Their primary lever of power is the AGM. Under the Australian Corporations Act, if more than 25% of votes are cast against the remuneration report (a "first strike"), and this is repeated the following year (a "second strike"), a spill motion is triggered and **all Board members must stand for re-election**.

4.4 Shareholder Advisers: e.g. Australian Shareholders Association

Shareholder advisory bodies such as the Australian Shareholders' Association (ASA) act as organised, values-driven intermediaries between listed companies and dispersed retail shareholders. Founded in 1960, ASA is Australia's largest independent, not-for-profit association representing individual investors, with a stated vision "to be the leading

independent voice and community for all Australian shareholders and investors." Its purpose is "to advocate for shareholders and create a better investment environment through investor engagement and education," and it describes itself as both a shareholder advocate and investor educator (Australian Shareholders' Association, 2023).

4.4.1 Values and Governance Focus

ASA's foundational values and governance documents emphasise transparency, independence, and equitable treatment of retail shareholders. Through its focus-issues and voting-guidelines program, ASA pushes listed companies to act "in the best interests of all shareholders, especially individual investors," with particular attention to fair capital raisings, effective shareholder engagement and AGM participation, board accountability and director standards, and transparent, performance-linked executive remuneration. ASA's advocacy and monitoring activities include written submissions to regulators and government, appearances at hearings, and direct engagement with company chairs and directors, giving it a structured channel to influence corporate governance norms in Australia (Australian Shareholders' Association, n.d.-b).

4.4.2 Operational Model: Company Monitors and Proxy Voting

Operationally, ASA relies on a network of volunteer company monitors who analyse annual reports, financial statements, and governance practices, meet with boards, and then prepare formal voting intentions for AGMs in accordance with ASA's published guidelines. ASA typically holds open proxies from its members and exercises them on key resolutions, publishing rationales that reflect its charter and values; it often becomes "the principal public questioner" at ASX-listed company meetings, particularly where there are many retail shareholders. This structure gives ASA a credible capacity to coordinate and signal the views of otherwise fragmented retail investors, and its recommendations can materially influence AGM outcomes in contentious remuneration or director re-election votes (Australian Shareholders' Association, n.d.-a).

4.4.3 Institutional Proxy Advisers

In the Australian market, ASA sits alongside institutional proxy advisers and governance research firms that primarily serve large superannuation funds and other wholesale investors. These organisations, such as Ownership Matters and other proxy advisers regulated under the Australian Financial Services Licence (AFSL) regime, produce independent governance research and voting recommendations, with regulatory settings designed to ensure independence and manage conflicts of interest. While their client base differs, both ASA and institutional proxy advisers play a similar functional role in our model: they synthesise complex governance information, apply explicit principles and guidelines, and publicly or semi-publicly recommend voting positions that other stakeholders then incorporate into their own decision processes (Ownership Matters, n.d.).

4.4.4 Modelling Implications

For modelling purposes, the ASA material maps to the adversarial framework through four steps:

(i) **Observed institutional features.** A governance charter centred on retail shareholder protection, a volunteer-monitor network that produces independent company assessments, and a track record of public activism on remuneration and board accountability.

(ii) **Implied objectives.** A utility function that places primary weight on fair treatment of retail shareholders, executive accountability, and transparent communication, with lower weight on short-term share-price performance.

(iii) **Implied likely actions.** In a crisis with clear governance failures, a high prior probability of issuing a negative voting recommendation ("vote against" on the remuneration report and potentially on director re-elections), amplified by coordination with institutional proxy advisers who share similar governance principles.

(iv) **Model entry point.** These features parameterise the ASA digital twin and the random-utility model for ASA recommendations in Section 6.3, where charter-derived weights determine the probability that ASA recommends escalation at each decision node.

These modelling implications draw on ASA's published focus issues and voting guidelines (Australian Shareholders' Association, 2025).

4.5 The Qantas Board

The Qantas Board is the focal corporate governance body in this case study and the decision-maker that has engaged us. During the period relevant to this paper, it comprised non-executive directors including Chairman Richard Goyder and other independent directors with backgrounds in finance, law, public policy, media, technology, and aviation, whose collective decisions shaped Qantas's response to mounting reputational and regulatory pressures. The Board is subject to statutory duties under the Corporations Act 2001 (Cth), including the duty of care and diligence in section 180, the duty to act in good faith in the best interests of the company and for a proper purpose in section 181, and obligations around continuous disclosure and financial reporting enforcement by ASIC, as well as governance expectations under the ASX Corporate Governance Principles and Recommendations.

4.5.1 Governance Review Findings

The Qantas Governance Review, commissioned in October 2023 and reported in August 2024, concluded that while there were "no findings of deliberate wrongdoing," "mistakes were made" by the Board and management and that "in some cases, the responses of the Board and Management were a contributing factor" to significant reputational and customer-service issues. The review identified five root-cause dynamics behind the loss of trust:

1. A leadership culture that at times prioritised financial performance over non-financial stakeholder risks (other than safety).

2. "Top-down leadership with a dominant and trusted CEO, leading to insufficient listening and low speak up."

3. A Board modus operandi that "did not always achieve the right balance between support and challenge."

4. Crisis-driven decision-making during and after COVID-19 that did not fully appreciate cumulative impacts on stakeholders.

5. External communications that "were at times combative which exacerbated issues."

Independent commentary on the review described it as finding that there had been "too much deference" to the long-tenured CEO and that the Board had failed to interrogate non-financial risks with the same intensity as financial metrics (Qantas Airways, 2024).

4.5.2 Escalation After the ACCC Proceedings

Against this backdrop, the announcement on 31 August 2023 that the ACCC had commenced Federal Court proceedings against Qantas over alleged "ghost flights" materially escalated the Board's governance and regulatory risk environment. The ACCC alleged that Qantas engaged in false, misleading or deceptive conduct by advertising and selling tickets for thousands of flights that had already been cancelled and by delaying notification to customers, signalling potential contraventions of the Australian Consumer Law and exposing the company to very substantial civil penalties and customer remediation. Combined with the High Court's later rejection of Qantas's appeal on the unlawful outsourcing of ground-handling staff, and with existing Senate inquiries, customer service failures, and executive remuneration controversy, the ACCC proceedings sharpened the Board's exposure across three liability channels:

1. Personal regulatory liability for individual directors under their statutory duties.

2. Board-level accountability through the two-strikes rule and spill mechanisms at the AGM.

3. Corporate-level exposure to class actions and regulatory penalties.

These liability channels were identified in contemporary analysis of the crisis (RMIT University, 2023).

4.5.3 Improvement Themes and Cultural Reset

The Governance Review explicitly linked these events to the need for a different Board culture and decision process, concluding that "the way we operate needs to change" and calling for "more robust debate and deeper examination of options and risks, particularly during times of volatility." It recommended ten "Improvement Themes," including more comprehensive risk identification and escalation (with greater weight on non-financial risks), changing the Board's modus operandi to "free up more time and attention for important matters and to challenge Management more effectively," and coalescing around new interaction norms focused on "sharing and collaboration, accountability and challenge, respect and support." Industry analysts interpreted these themes as a direct response to what they described as an overly "submissive" or deferential Board dynamic that had allowed a dominant CEO and a narrow focus on financial performance to persist despite accumulating legal, regulatory, and reputational warning signs (Australian Aviation, 2024).

4.5.4 Modelling Implications

For modelling purposes, the Board evidence maps to the adversarial framework through the same four-step template:

(i) **Observed institutional features.** Statutory duties under the Corporations Act 2001 (care, diligence, good faith), ASX Corporate Governance Principles compliance, a governance review finding of excessive deference to a dominant CEO, and escalating regulatory and litigation exposure following the ACCC proceedings.

(ii) **Implied objectives.** A utility function with evolving weights: prior to the ACCC crisis, higher weight on financial performance and CEO continuity; after the crisis, sharply increased weight on legal-liability mitigation, stakeholder trust restoration, and long-term governance credibility, reflecting the shift from a "supportive and deferential" to a "challenge and accountability" Board culture.

(iii) **Implied likely actions.** A rising probability over time of commissioning an independent governance review, adjusting or clawing back executive remuneration, and initiating CEO transition, driven by the combination of personal director liability, two-strikes risk at the AGM, and reputational contagion from ongoing regulatory proceedings.

(iv) **Model entry point.** These features parameterise the Board's decision probabilities at each game-tree node in Section 6.2, where the utility weights shift at the ACCC-announcement checkpoint and the argmax-count decision rule converts weighted expected utilities into action probabilities.

The Board's governance obligations are framed by the ASX Corporate Governance Principles and Recommendations (ASX Corporate Governance Council, 2024).

4.6 Regulators

Although regulators are not modelled as strategic decision-makers in the game tree (their actions are exogenous to the Board's decision horizon), the regulatory environment is a critical contextual factor that shapes every stakeholder's behaviour. Three regulatory channels are relevant to the Qantas case.

4.6.1 Workplace Regulation: Fair Work Ombudsman

In March 2020, the Fair Work Ombudsman (FWO) accepted a court-enforceable undertaking from Qantas after the airline self-reported years of underpayments for some marketing and administrative staff. The undertaking required multi-million-dollar back pay with interest, public apologies, and independent audits to verify future compliance. The FWO emphasised that the action was both remedial and deterrent: rectifying underpayments, improving Qantas's internal payroll systems, and "levelling the playing field" for law-abiding employers. The use of an enforceable undertaking rather than immediate litigation reflected a pragmatic regulatory strategy, securing concrete outcomes while conserving enforcement resources.

4.6.2 Competition and Consumer Protection: ACCC

The ACCC launched Federal Court proceedings on 30 August 2023 alleging that Qantas had engaged in false, misleading, or deceptive conduct by advertising and selling tickets for more than 8,000 flights it had already cancelled and by delaying notification to existing ticket-holders for thousands more. The ACCC's stated goal was to protect consumers and deter similar conduct, arguing that Qantas's behaviour left customers with less time to re-plan travel and may have forced them to pay higher prices for alternative flights. The Commission signalled it would seek what was reported as a record penalty, understood to be in excess of A$250 million, to achieve both specific and general deterrence.

In this domain, the ACCC has direct investigative powers (compulsory information notices, data analysis) but relies heavily on the success of court proceedings under the Australian Consumer Law to obtain penalties, injunctions, and compensation orders. The ultimate impact of its action against Qantas therefore depends on judicial findings rather than unilateral administrative fines.

4.6.3 Workplace Enforcement: Union-Led Litigation

Workplace enforcement in the outsourcing and mass-redundancy disputes primarily turned on private and union-led litigation under the Fair Work Act rather than administrative investigation by a regulator such as the FWO. The Transport Workers' Union's challenge to the 2020 outsourcing culminated in the High Court's unanimous confirmation in September 2023 that Qantas had unlawfully outsourced approximately 1,700 ground-handling roles. In this context, the Fair Work framework gives regulators tools such as enforceable undertakings and, in some circumstances, civil penalty proceedings, but the key remedies (declarations of illegality, penalties, and compensation) depend on the courts' interpretation and application of the statute, making court success central to meaningful sanctions and deterrence.

4.6.4 Regulatory Power and Its Limits

Across both labour and consumer domains, regulators' direct powers (investigations, undertakings, public enforcement announcements) are important for shaping behaviour and public expectations, but large monetary penalties and binding orders in the Qantas matters have largely required successful litigation or court-endorsed settlements rather than purely administrative action. This distinction matters for the game tree: the Board's calculus around regulatory risk is driven not by the certainty of a fixed administrative fine but by the *probability distribution* over judicial outcomes, a distribution that the ARA framework is well-suited to model.

4.7 Stakeholder Summary

Table 1 consolidates the stakeholder architecture before the modelling sections begin.

Table 1: Stakeholders, power channels, modelling roles, and primary evidence sources. For the statistical model form and decision rule associated with each strategic actor, see Table 29.

Stakeholder	Why they matter	Direct lever	Strategic actor?	Model form	Main evidence
Board	Focal decision-maker	Direct: review, reset, remuneration	Yes	Utility + ordinal probit + recursion	Governance review, legal context
ASA	Governance activist / signaller	Direct: voting recommendations	Yes	Random utility / Beta integration	Charter, focus issues, voting practice
Shareholders	AGM vote	Direct: remuneration vote	Yes	Logit-normal vote model	Historical AGM voting panel
CEO	Resistance / resignation	Direct: compliance or defiance	Yes	CRRA + opponent priors	Governance review, public commentary
Customers	Background reputational pressure	Indirect: media, sentiment	No	Contextual only	Market share, reputation metrics
Regulators	Context shaper	Indirect over horizon	No	Exogenous context	ACCC, FWO, litigation record

5 Stakeholder Simulation via Generative AI

Generative AI is not used here because it is fashionable; it is used because the problem would otherwise be intractable. The key stakeholders are inaccessible, the event is hindsight-contaminated, the historically comparable sample is sparse, and the model requires structured responses across many counterfactual scenarios. Digital twins provide a scalable, auditable way to generate scenario-specific preference information that can then be constrained by historical data and governance logic.

5.1 Digital Twins

In this case study, digital twins are the central mechanism for bringing generative AI into adversarial risk analysis. We define a digital twin as a generative-AI agent calibrated to the public behaviour, stated values, and revealed preferences of a real stakeholder, such as the Qantas Board, the CEO, or the Australian Shareholders' Association. These agents are implemented using large language models and are queried under carefully designed scenarios that mirror real governance decisions, allowing us to elicit structured judgements about how those stakeholders are likely to react.

Digital twins serve two distinct roles in this paper. First, **preference elicitation at scale**: they allow us to elicit preferences across many scenarios that would be impractical to test with real individuals or institutions, including "what-if" combinations

that never occurred historically. The Board digital twin is queried across 95 structured governance scenarios (Section 6.2.2), each repeated 40 times with randomised factor orderings, producing 3,800 observations. This volume of preference data would be impossible to obtain from real directors (no board would sit through 95 hypothetical governance scenarios), but is essential for identifying the 10+ utility weight parameters in the Board's utility function. Second, **calibration anchoring**: they provide an internally consistent, reproducible set of responses that we use as an anchoring prior when historical data are sparse or noisy, particularly for reputational events where observed outcomes reflect both behaviour and exogenous noise. The ASA digital twin provides range estimates, floor probabilities, and pairwise gap assessments that anchor the constrained optimisation pipeline for ASA action probabilities. Again, the ASA would not provide these assessments directly, but the LLM's training on ASA voting records, policy statements, and media commentary allows it to produce credible approximations. In both roles, the digital twin is not treated as ground truth; it is an information source that complements, and is constrained by, empirical data, governance theory, and expert judgement. The digital twin's outputs are *data* that enter a Bayesian estimation pipeline, where they are combined with informative priors, subjected to convergence diagnostics, and validated against out-of-sample historical outcomes. The LLM is an instrument, not an oracle.

More generally, large language models provide a transformative capability for adversarial risk analysis in reputation-risk settings. Digital twins allow the analyst to:

- Elicit preferences at scale and across scenarios that would be impractical to test with real individuals.
- Stress-test models by simulating stakeholder responses to novel or extreme events.
- Fill data gaps where historical precedent is sparse or where direct access to the decision-maker is unavailable.
- Provide an auditable, reproducible record of the assumptions underlying each stakeholder model.

Generative AI does not replace actuarial judgement; it extends the reach of actuarial methods into domains that were previously data-poor and model-resistant, with reputation risk as a leading example.

5.1.1 Research Support

There is now a growing body of research on using LLMs to simulate human judgement and behaviour. Aher, Arriaga, and Kalai (2023) show that large language models can simulate multiple humans and replicate the main findings of human-subject studies across a range of experimental designs, including social dilemmas and policy attitudes. Argyle et al. (2023) use language models to generate simulated human samples and find that, when carefully conditioned, model outputs closely match survey-based distributions on political and social questions. Related work on "generative agents" (Park et al., 2023) demonstrates that LLM-based agents embedded in a structured environment can exhibit human-like patterns of planning, interaction, and memory over time. At the same time, this literature warns that LLMs inherit training-data biases, may under-represent minority perspectives, and can fail in fine-grained or high-stakes prediction tasks; recent overviews explicitly

caution against treating LLM simulations as direct substitutes for human subjects or as authoritative forecasts (Binz & Schulz, 2023). Our use of digital twins reflects that caution: they are treated as noisy, scenario-specific priors about stakeholder tendencies, not as definitive predictions about any particular director, executive, or institution.

5.1.2 Origins of the "Digital Twin" Concept

The term "digital twin" itself comes from engineering and product lifecycle management rather than AI. Grieves and Vickers introduced it to describe a virtual representation of a physical product or system that is continuously updated with sensor data from the real world, allowing engineers to monitor performance, test interventions, and optimise maintenance strategies on the virtual model before acting on the physical asset (Grieves & Vickers, 2017). In aerospace, for example, a digital twin of a jet engine tracks operating conditions, wear, and failure modes, providing a high-fidelity simulation that is tightly coupled to the real engine. Our use is analogical: instead of an engine, we create digital twins of stakeholder groups whose configuration and constraints are informed by charters, legal duties, historical decisions, and public communications, and then subject those twins to counterfactual events (ACCC actions, first-strike votes, governance reviews) to estimate plausible reaction patterns.

5.1.3 Elicitation and Calibration in Practice

Constructing useful digital twins requires more than asking an LLM "what will the Board do?"; it depends critically on prompt engineering and on a clear separation between elicitation and calibration roles. For **preference elicitation**, we design prompts that put the model in the position of the stakeholder, specify the decision problem, and require explicit trade-off reasoning. For example, the Board twin is instructed to simulate a 2023 boardroom discussion, with named directors and a structured agenda, and to weigh three liability channels (personal ASIC exposure, two-strikes spill risk, and corporate legal penalties) alongside reputation and capital considerations. The prompt requires the twin to output probabilities over discrete governance actions and to explain how internal disagreements are resolved, which we then use as inputs to an ordinal-probit model for Board decisions.

For **calibration anchoring**, we use digital-twin responses as a structured prior that is reconciled with historical data rather than overriding it. For example, the ASA twin is calibrated using its published charter, focus issues, and voting guidelines, and then asked to evaluate hypothetical remuneration-report scenarios that resemble, but are not identical to, past Qantas and non-Qantas cases. The resulting pattern of "vote for / vote against / abstain" recommendations serves as an anchor for the parameters of a random-utility model, which is then constrained by observed ASA voting recommendations in historical AGMs. Similarly, CEO and shareholder twins are used to generate prior beliefs about sensitivity to reputation shocks and governance actions, but posterior distributions are dominated by empirical event studies and observed vote-outcomes where data are available.

Across all stakeholders, we follow a consistent goal–context–few-shot design. Each twin is given (i) a clear role and objective, (ii) rich contextual information (legal duties, institutional constraints, historical events, peer cases), and (iii) carefully chosen few-shot examples that illustrate the desired style of deliberation and the required output schema (factor ratings, probability vectors, narrative commentary). This approach is consistent

with the broader prompt-engineering literature (Wei et al., 2022; Brown et al., 2020), which finds that instruction prompts, role assignment, and in-context examples materially improve the reliability and face validity of LLM-based behavioural simulations. Within our framework, however, digital twins remain tools for **elicitation** and **calibration anchoring**: their outputs are always interpreted probabilistically, checked against historical Qantas events where possible, and embedded in transparent Bayesian models rather than being treated as ground-truth predictions.

5.2 Prior Work: Literature Review

The use of LLMs as simulated human participants has emerged as a distinct research programme since 2022, with contributions from computational social science, economics, and cognitive psychology.

Horton (2023) introduced the concept of *Homo Silicus*, using LLMs as simulated economic agents to replicate classic experimental results in behavioural economics. His experiments showed that GPT-3.5 and GPT-4 reproduce well-known phenomena including ultimatum game behaviour, dictator game allocations, and labour supply responses, with effect sizes comparable to human subject pools. The key methodological insight is that LLMs can serve as a "computational wind tunnel" for testing economic theories before running expensive human experiments.

Argyle et al. (2023) demonstrated that LLMs can simulate demographically conditioned human survey responses with surprising accuracy. By conditioning GPT-3 on demographic backstories (age, race, education, political affiliation), they generated synthetic survey responses that reproduced known subgroup differences in political attitudes, voting intentions, and policy preferences. Their "silicon sampling" approach showed correlations of $r = 0.85$–0.95 with actual survey data from the American National Election Studies.

Aher et al. (2023) replicated 10 classic social science experiments using GPT-4 as both experimenter and subject pool, including the Milgram obedience study, the Ultimatum Game, and the Garden Path sentence processing task. They found that GPT-4 reproduced the qualitative patterns of human behaviour in 8 of 10 studies, though with some systematic biases (e.g., greater cooperativeness than human participants).

Park et al. (2023) created persistent generative agents (LLM-driven characters with memory, reflection, and planning capabilities) that exhibited emergent social behaviours including information diffusion, relationship formation, and coordination for group activities. Their architecture demonstrated that LLMs can sustain coherent persona-consistent behaviour over extended interactions, a property essential for digital twin applications where the same stakeholder must respond consistently across dozens of scenarios.

In the domain of corporate and market research, Brand, Israeli, and Ngwe (2023) showed that GPT-4 can predict consumer preferences, willingness to pay, and brand perceptions with accuracy comparable to traditional conjoint analysis, at a fraction of the cost. Shanahan, McDonell, and Reynolds (2023) provided a theoretical framework for understanding LLM role-play as a form of conditional text generation, arguing that the quality of simulation depends critically on the specificity and coherence of the role description.

Binz and Schulz (2023) systematically evaluated GPT-3's performance on cognitive psychology tasks, finding that it matched human decision-making patterns on tasks

involving risk, uncertainty, and intertemporal choice, precisely the cognitive domains relevant to board governance decisions. However, they also identified systematic deviations: GPT-3 was more risk-neutral than typical human subjects and less susceptible to framing effects, suggesting that prompt engineering must actively introduce the biases that real decision-makers exhibit.

This literature establishes that LLM-based stakeholder simulation is methodologically viable but requires careful calibration. The present case study extends this work in three directions: (1) applying digital twins to a specific real-world governance crisis rather than stylised experimental settings; (2) embedding the LLM outputs within a formal Bayesian estimation pipeline rather than treating them as point predictions; and (3) using structured prompt engineering to control for known cognitive biases rather than relying on the LLM's default behaviour.

5.3 Prompt Engineering

The quality of a digital twin depends critically on the prompt architecture: the system instructions, persona definition, contextual framing, and response format that shape the LLM's simulated behaviour. Poor prompts produce generic, mode-seeking responses; well-engineered prompts produce calibrated, scenario-sensitive distributions.

The elicitation system uses a two-part prompt architecture:

1. **System prompt**: fixed across all queries. Establishes the stakeholder persona, legal context, historical context, and response format. For the Board digital twin, this names all eight directors explicitly (Chairman Richard Goyder and seven independent directors), specifies their professional backgrounds, and instructs the LLM to simulate a boardroom deliberation where directors raise competing concerns and work toward a majority position.

2. **Scenario prompt**: varies per query. Describes the specific game state (CEO status, vote outcome, prior Board actions, review findings) and asks for action probabilities and factor ratings.

Three design principles govern the prompt engineering:

Persona grounding. The system prompt anchors the simulation in specific, named individuals rather than generic archetypes. Research on LLM role-play (Shanahan et al., 2023) shows that named characters with detailed backstories produce more consistent and differentiated responses than abstract roles. The Board prompt names all eight directors and their professional domains (finance, law, public policy, media, technology, aviation operations), enabling the LLM to generate internally diverse deliberations rather than monolithic responses.

Cognitive bias framing. A naive LLM, prompted as a corporate board without further context, produces unrealistically moderate responses, assigning 25–35% probability to CEO transition even in severe crisis scenarios, compared to the empirical base rate of 100% in comparable ASX 100 ESG crises. This occurs because the LLM's default behaviour reflects the average of its training distribution, which includes many non-crisis governance decisions where CEO retention is the norm.

The prompt corrects this by explicitly activating four bias mechanisms that real directors experience during a severe governance crisis:

- **Anchoring** (Tversky & Kahneman, 1974): stating the empirical base rate ("100% of comparable cases resulted in CEO departure") as a numerical anchor.
- **Bandwagon/social proof** (Westphal & Zajac, 1997): naming five specific peer companies (AMP, Crown Resorts, Rio Tinto, Westpac, NAB) and their crisis-driven CEO departures.
- **Loss aversion** (Kahneman & Tversky, 1979): framing inaction as an active choice with personal regulatory consequences for directors under ASIC s 180 enforcement.
- **Counter-bias deactivation** (Fischhoff, 1982; Larrick, 2004): explicitly naming five cognitive biases that favour CEO retention (escalation of commitment, status quo bias, groupthink, over-optimism, hyperbolic discounting) so the LLM recognises and discounts them.

These are not manipulative tricks: they represent genuine information available to real directors during a severe governance crisis. The prompt makes this information explicit rather than relying on the LLM's implicit training distribution. The result is a calibrated shift: with all bias mechanisms active, the Board digital twin assigns 90–97% probability to CEO transition in severe crisis scenarios, matching the empirical base rate.

Structured output with randomised controls. Each query returns a JSON-validated response containing action probabilities (summing to 1.0), ten factor severity ratings on a 1–5 Likert scale, and free-form deliberation commentary. The ten factors are presented in **randomised order** across the 40 repetitions per scenario, serving as a diagnostic for anchoring bias: if factor presentation order systematically affects ratings, the LLM is exhibiting order-dependent behaviour that must be corrected. The Pydantic validation schema enforces probability normalisation (within 0.02 tolerance, auto-renormalised) and factor completeness (all 10 indices present after deduplication).

5.4 Temperature and Repeated Sampling

A single LLM call produces a point estimate: a single probability vector over actions. This is insufficient for Bayesian estimation, which requires a *distribution* of responses that captures the LLM's own uncertainty. Two mechanisms generate this distributional data:

Temperature sampling. The LLM is called with temperature $T = 1.0$ (the model's default stochastic sampling mode). At $T = 1.0$, the softmax over the model's logits is unmodified, producing diverse but coherent responses. Lower temperatures ($T < 0.5$) would concentrate outputs near the mode, reducing diversity and producing artificially tight posterior estimates. Higher temperatures ($T > 1.5$) would introduce incoherent responses that fail validation. The choice of $T = 1.0$ balances diversity against coherence.

Temperature sampling generates **aleatoric-like** variation: each call produces a different draw from the LLM's conditional distribution over responses, analogous to sampling different "virtual directors" from the population of boards consistent with the prompt.

Repeated sampling. Each scenario is queried $n = 40$ times (seeds 0–39), each with an independently randomised factor presentation order. The 40 repetitions serve three purposes:

1. **Distributional estimation.** The mean and variance of the 40 probability vectors provide the point estimate and uncertainty measure that enter the downstream Bayesian pipeline. For the Board, the ordinal probit model uses the individual Likert ratings from all 3,800 observations. For the ASA, the mean and standard deviation of $P(\text{strike})$ across 30 draws provide the moments for method-of-moments Beta fitting.

2. **Order-effect diagnostics.** Comparing factor ratings across different presentation orders tests whether the LLM exhibits anchoring bias. Systematic order effects would indicate that the ratings are an artefact of presentation rather than a reflection of the scenario's governance characteristics.

3. **Standard error reduction.** Quantitative analysis showed a 29% standard error reduction per doubling of repetitions. At $n = 40$, the widest confidence interval across all utility weight parameters narrows sufficiently for meaningful inference. The marginal cost is modest: approximately \$0.60 per full run of 95 scenarios at gpt-4o-mini pricing.

The combination of temperature sampling and repeated querying transforms the LLM from a deterministic predictor into a **stochastic data source** whose outputs can be treated as noisy observations of an underlying preference structure, exactly the input that a Bayesian estimation pipeline requires.

Seed determinism and reproducibility. Each LLM call is seeded by `SHA-256 (scenario_prompt + seed)`, ensuring deterministic factor ordering across processes (unlike Python's `hash()`, which is process-randomised via `PYTHONHASHSEED`). All responses are cached using SHA-256 keys computed from the scenario prompt, model name, seed, and cache version, enabling exact reproducibility of the full elicitation dataset.

Table 2: Digital twin workflow used in this paper.

Stakeholder	Source material	Role of twin	Output type	How anchored to data	Model section
Board	Governance review, legal context, public signals	Preference elicitation	Action probabilities and rationale	Ordinal probit on 3,800 scenario responses	6.2
ASA	Charter, focus issues, voting guidelines	Calibration anchor	Vote/support ranges and floor estimates	Constrained optimisation + historical recommendations	6.3
CEO	Governance review, public commentary, crisis behaviour	Opponent prior	Resignation / resistance tendencies	Beta posterior from 12/12 historical departures	6.5

6 Stochastic Modelling Process

6.1 Decision Tree

We model this situation as a **turn-taking game with a sequence of decisions**. The game tree is built up progressively through four stages, each adding a new layer of strategic interaction. In the figures that follow, nodes are colour-coded by actor (red for CEO decisions, blue for Board decisions, green for ASA decisions, and grey for Nature (chance) nodes), with diamond-shaped terminal nodes showing expected utilities. The **red line** traces the path corresponding to the actual outcomes observed in 2023–2024.

6.1.1 Stage 1: The CEO's Pre-Game Decision

The tree begins with the CEO's pre-game decision node D_0^{ceo} (Figure 4). The CEO must choose whether to resign or stay. The model assigns a 96.2% probability to resignation and 3.8% to staying, consistent with the empirical Beta(12.5, 0.5) prior from 12 no-contrition CEO departures (Table 21) refined by Level-2 ARA analysis. The popup in the figure shows the expected utilities for each action, confirming that resignation dominates. The red line follows the actual outcome: Joyce resigned on 5 September 2023.

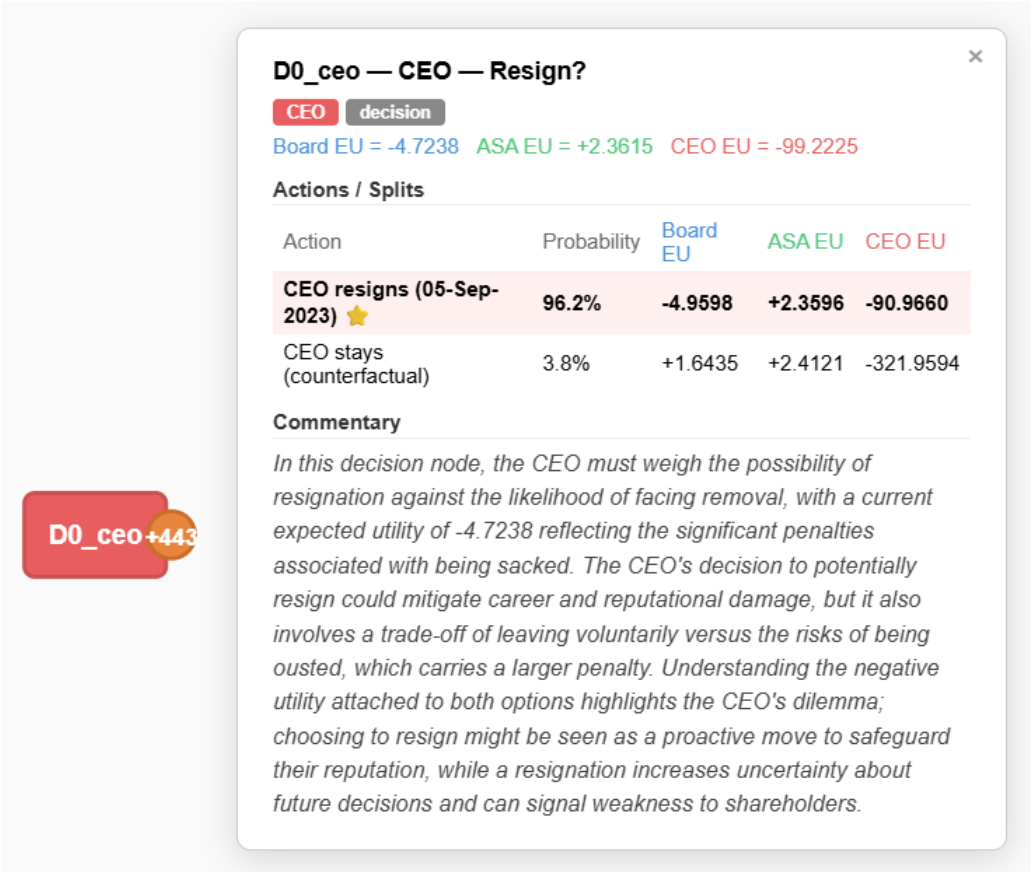

Figure 4: Stage 1: The CEO's pre-game decision node (D_0^{ceo}). The model predicts resignation with 96.2% probability. The popup shows expected utilities and commentary for each action. The red line marks the actual outcome.

6.1.2 Stage 2: The Board's Initial Response

Given the CEO's choice, the tree expands to the Board's first decision node D_1 (Figure 5). Both branches (CEO resigns (red line) and CEO stays (counterfactual)) lead to a Board decision among three options: do nothing (D0_minimal), commission a governance review (D1_review), or force CEO transition (D3_ceo_transition). The expected utility badges on each D_1 node show that commissioning a review dominates in both branches, though with different magnitudes. The red line follows the actual path: CEO resigned, Board chose to commission a review.

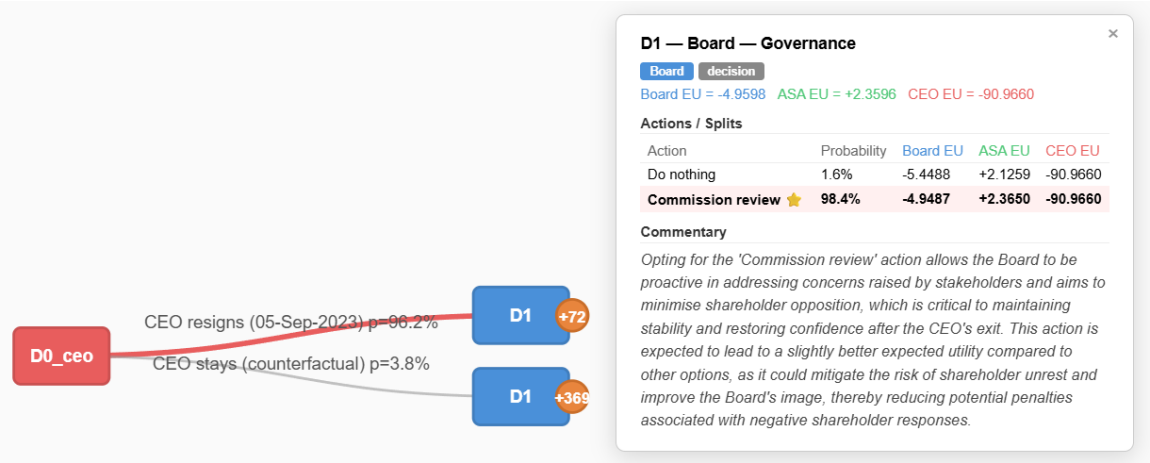

Figure 5: Stage 2: The Board's initial response (D_1). Both the CEO-resigns path (red line) and CEO-stays counterfactual lead to Board decision nodes. Expected utility badges show the review option dominates.

6.1.3 Stage 3: ASA, Shareholders, and Post-AGM Decisions

The tree now expands substantially (Figure 6) to include the ASA's strike recommendation (A_2), the shareholder vote (Nature node V), the Board's post-AGM response (D_{rev}), and the governance review outcome (Nature node R). Each Board action at D_1 fans out into ASA recommendation branches (with strike probabilities exceeding 96% across all nodes), which in turn fan out into vote outcomes classified as no-strike, first-strike, or overwhelming. The post-AGM Board response and review findings produce the first terminal nodes (grey diamonds). The red line traces the actual path: CEO resigned → Board commissioned review → ASA recommended strike → overwhelming vote (82.9%) → review finding of "balanced" → terminal outcome.

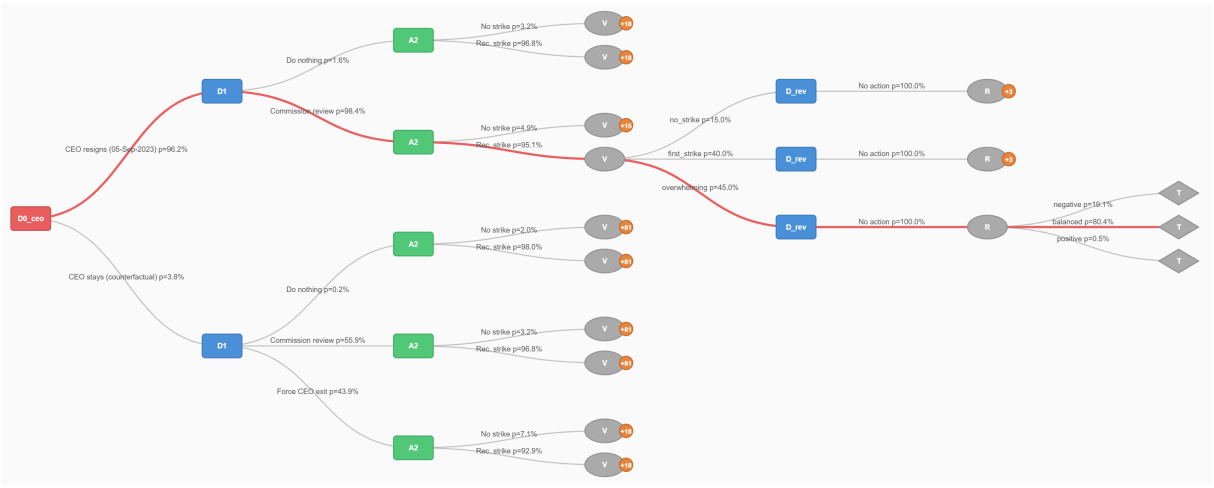

Figure 6: Stage 3: The tree expands through ASA decisions (A_2, green), shareholder votes (V, grey), Board post-AGM responses (D_{rev}, blue), and review outcomes (R, grey diamonds). The red line traces the actual 2023–2024 sequence through to the balanced review finding.

6.1.4 Stage 4: Post-Review CEO Decisions and Full Terminal Nodes

The final expansion (Figure 7) adds the CEO's post-AGM decision nodes (D_4 and D'_4) in the counterfactual branch where the CEO stayed, plus post-review Board and CEO responses (D'_{rev}, $D_4^{\text{post-review}}$). In the CEO-stayed branch, after an adverse review finding the CEO faces a further decision: stay, resign late, or negotiate an exit. The Board may also act post-review by sacking the CEO. These additional layers produce the full set of terminal nodes, each carrying a complete game-state description (CEO status, vote outcome, review finding, Board actions taken) from which expected utilities are computed.

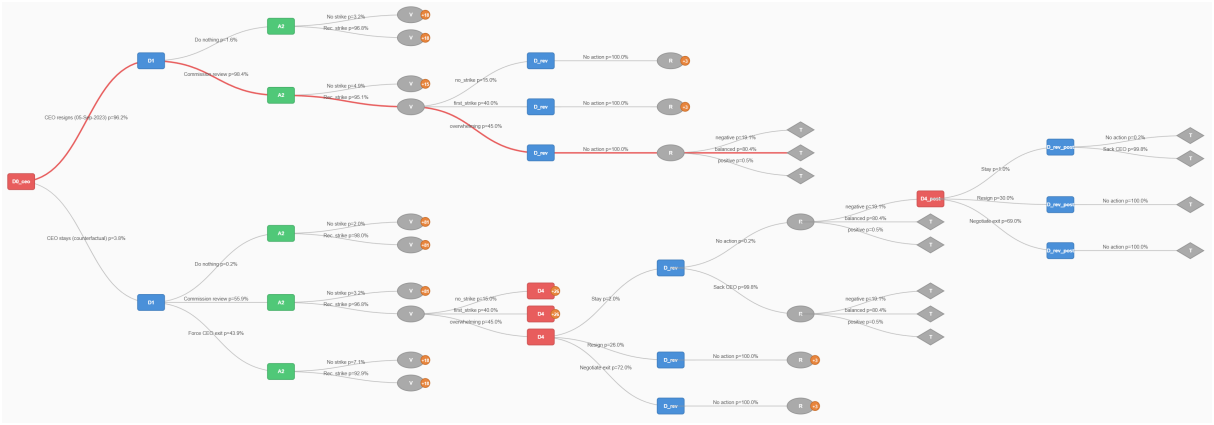

Figure 7: Stage 4: Further expanded tree with post-review CEO decisions (D'_4, red) and post-review Board responses (D'_{rev}, blue) in the CEO-stayed counterfactual. More branches now terminate in outcome nodes with computed expected utilities. The red line continues to trace the actual outcome path.

6.1.5 The Fully Expanded Tree

When all decision nodes, chance nodes, and conditional branches are expanded to their full depth (including per-draw Dirichlet noise on action probabilities, Monte Carlo vote samples, and nested ARA predictive distributions), the resulting tree is considerably more complex than the schematic stage-by-stage views in Figures 4–7. Figure 8 is included as an existence proof of this complexity rather than as a figure to be read node-by-node. The tree contains several hundred nodes across up to eight sequential decision layers, reflecting the combinatorial expansion of multi-step multi-adversary games that Ekin et al. (2021) identified as the central computational challenge for ARA. In practice, the engine traverses this tree via backward induction with Monte Carlo integration at each chance node, using the computational budget described in the subsections that follow.

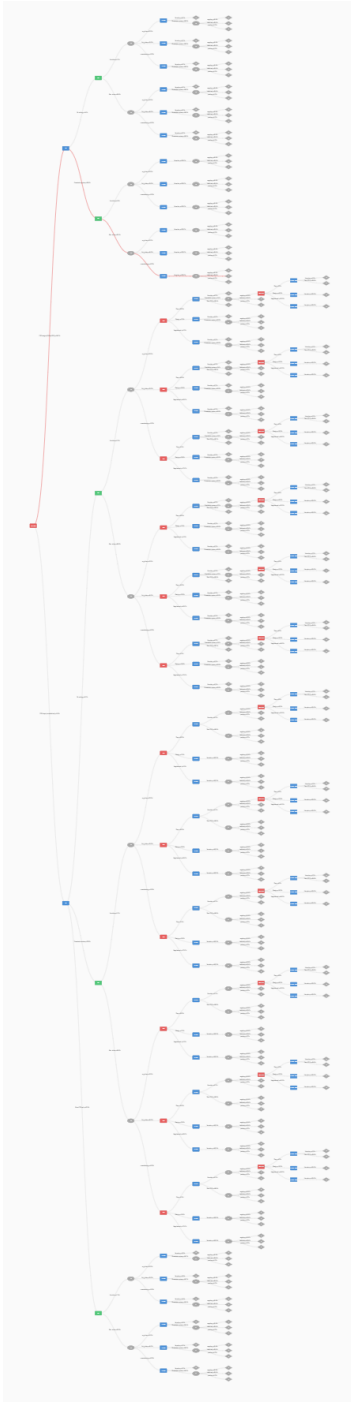

Figure 8: The fully expanded game tree contains several hundred nodes across eight decision layers; this figure is included to illustrate structural complexity rather than to be read node-by-node. The progressive stage-by-stage views in Figures 4–7 provide the readable decomposition.

6.1.6 Summary of Decision Sequence

The key decisions, in approximate chronological order, are:

1. **The CEO** (D_0^{ceo}) decides whether to resign voluntarily or remain in position.
2. **The Board** (D_1) decides whether to:

- Take no pre-emptive action (status quo).
- Commission an independent governance review.
- Force CEO transition.

3. **The Australian Shareholders Association** (A_2) decides whether to recommend voting against the remuneration report, recommend voting in favour, or make no recommendation.

4. **Shareholders** (V, Nature node) vote on the remuneration report, producing a continuous vote fraction from which strike and overwhelming indicators are derived.

5. **The Board** (D_{rev}) decides on a post-AGM response: no action, commission review, or sack CEO.

6. **The CEO** (D_4, D'_4, if still present) decides whether to stay, resign, or negotiate an exit.

7. **The Review** (R, Nature node) resolves with a qualitative outcome (negative, balanced, positive), a cumulative abnormal return, and direct costs.

6.2 Modelling the Board

The Board is our client and the primary decision-maker. We model their decisions through a three-stage pipeline: (1) scenario elicitation using LLM digital twins, (2) Bayesian estimation of utility weights via an ordinal probit model, and (3) stochastic simulation of decisions via argmax-count over posterior draws.

6.2.1 The Board Utility Function

The Board's expected utility for a terminal outcome Z is decomposed into a weighted sum of **basis functions** ϕ_k that capture distinct governance concerns, plus anchored terms for market and cost impacts:

$$\text{EU}(Z) = \sum_{k=1}^{K} w_k \cdot \phi_k(Z) + \text{anchored}(Z) \tag{1}$$

where $w_k > 0$ are the utility weights to be estimated and $\phi_k(Z)$ are indicator or graduated basis functions computed from the game state. The anchored term captures cumulative abnormal return (CAR) and direct review costs scaled by fixed anchors ($w_{\text{CAR}} = 15.0$, $w_{\text{cost}} = 15.0$).

The full utility function has three structural layers. Define:

- $\text{ceo\_at\_end} = \text{not CEO\_removed} \wedge \neg\text{CEO\_resigned\_early}$
- $\text{removed\_involuntary} = \text{CEO\_removed} \wedge \neg\text{CEO\_resigned\_early}$
- $\text{board\_inactive} = (d_1 = \text{D0\_minimal}) \wedge d_{\text{rev}} \notin \{\text{sack}, \text{review}\} \wedge d'_{\text{rev}} \neq \text{sack}$

$$
\begin{aligned}
u_B(Z) = & \underbrace{-w_{\text{inact\_base}} \cdot \mathbf{1}[\text{board\_inactive}] - w_{\text{inact\_no\_rev}} \cdot \mathbf{1}[\text{not review\_comm}]}_{\text{1a. Inaction (unconditional)}} \\
& \underbrace{-w_{\text{inact\_ceo}} \cdot \mathbf{1}[\text{ceo\_at\_end}] - w_{\text{inact\_no\_sack}} \cdot \mathbf{1}[\text{not removed\_inv}]}_{\text{1b. Inaction (continued)}} \\
& \underbrace{-w_{\text{v\_strike}} \cdot \frac{(V - 0.25)_+}{0.75} - w_{\text{v\_over}} \cdot \frac{(V - 0.50)_+}{0.50}}_{\text{2. Vote penalties (linear in normalised excess)}} \\
& - w_{\text{pass}} \cdot \mathbf{1}[\text{CEO\_resigned\_early}] \\
& + w^+_{\text{CAR}} \cdot (\text{CAR})_+ \cdot \mathbf{1}[\text{review\_comm}] - w^-_{\text{CAR}} \cdot (-\text{CAR})_+ \cdot \mathbf{1}[\text{review\_comm}] \\
& - w_{\text{cost}} \cdot C_{\text{direct}} \cdot \mathbf{1}[\text{review\_comm}] \\
& - w_{\text{impl}} \cdot \big(\mathbf{1}[d_1 = \text{D3}] + \mathbf{1}[d_{\text{rev}} = \text{sack}] + \mathbf{1}[d'_{\text{rev}} = \text{sack}]\big) \\
& - \max\!\big(0,\ w_{\text{loss}} - w_{\text{loss\_over}} \cdot \mathbf{1}[\text{overwhelming}]\big) \cdot \mathbf{1}[\text{removed\_inv}] \\
& - w_{\text{rev\_neg}} \cdot \mathbf{1}[\text{review\_comm} \wedge \text{outcome} = \text{negative}] \\
& - w_{\text{rev\_bal}} \cdot \mathbf{1}[\text{review\_comm} \wedge \text{outcome} = \text{balanced}] \\
& - w_{\text{rev\_post}} \cdot \mathbf{1}[\text{removed\_inv} \wedge \text{not review\_comm}]
\end{aligned}
\tag{2}
$$

The review CAR term incorporates loss aversion following Kahneman and Tversky (1979):

$$
w^+_{\text{CAR}} = \frac{w_{\text{CAR}}}{\frac{1+\lambda_{\text{la}}}{2}}, \qquad w^-_{\text{CAR}} = \lambda_{\text{la}} \cdot w^+_{\text{CAR}} \tag{3}
$$

where $\lambda_{\text{la}} = 2.25$. Positive CARs receive weight ≈ 9.23; negative CARs receive weight ≈ 20.77, reflecting the empirical finding that boards weigh losses roughly twice as heavily as equivalent gains. The original Kahneman–Tversky estimate of $\lambda \approx 2.25$ was derived from individual monetary gambles rather than corporate market reactions; however, earlier versions of this model treated loss aversion as a free parameter estimated from the digital twin data, and the fitted value converged close to 2.25. On this basis, λ_{la} was fixed at the canonical value to reduce computational complexity. This convergence is itself noteworthy: it suggests that generic LLMs have learned to replicate loss-aversion patterns from their training corpus, lending additional support to the use of digital twins as calibration instruments.

The action-varying utility weights are estimated from the 3,800 digital-twin responses via the Stan ordinal-probit model described in Section 6.2.2. Posterior means and 95% credible intervals are shown in Table 3. Three further weights are not free parameters but fixed anchors: the review CAR weight $w_{\text{CAR}} = 15.0$ and direct-cost weight $w_{\text{cost}} = 15.0$ set the utility scale, and the loss-aversion coefficient $\lambda_{\text{la}} = 2.25$ is held at the Kahneman–Tversky canonical value (earlier fits converged near 2.25; see discussion above).

Parameter	Symbol	Posterior mean	95% CI
Inaction base penalty	$w_{\text{inact\_base}}$	6.14	[5.28, 7.20]
Inaction no review penalty	$w_{\text{inact\_no\_rev}}$	6.11	[5.70, 6.54]
Inaction delay penalty	$w_{\text{inact\_delay}}$	0.29	[0.06, 0.72]
Board passivity after departure	w_{pass}	2.18	[1.39, 2.97]
CEO removal cost	w_{removal}	0.02	[0.01, 0.03]
CEO removal shock (overwhelming)	$w_{\text{loss\_over}}$	0.01	[0.004, 0.02]
Negative review finding penalty	$w_{\text{rev\_neg}}$	4.07	[2.64, 5.51]
Balanced review finding penalty	$w_{\text{rev\_bal}}$	3.36	[0.95, 5.92]
Review after removal penalty	$w_{\text{rev\_post}}$	0.006	[0.002, 0.010]
CEO accountability benefit	w_{acct}	5.57	[5.45, 5.70]
Vote strike penalty	$w_{\text{v\_strike}}$	0.79	[0.16, 2.03]
Vote overwhelming penalty	$w_{\text{v\_over}}$	0.82	[0.17, 2.06]

Table 3: Board utility weight parameters: posterior means and 95% credible intervals from the Stan ordinal-probit fit to 3,800 digital-twin responses. Weights with tight CIs away from zero (inaction penalties, passivity, accountability, review findings) drive Board behaviour; weights with CIs close to zero (removal cost, review-post-removal) are effectively inactive in the fitted model. Fixed anchors (w_{CAR}, w_{cost}, λ_{la}) are discussed in the text.

6.2.2 Scenario Elicitation via LLM Digital Twins

There is no historical dataset of Australian board governance decisions with observable utility weights. The default parameters in Table 3 were set by expert judgement, but their magnitudes are uncertain and potentially inconsistent. We therefore use an LLM (specifically gpt-4o-mini) as a **calibrated stakeholder simulator** to generate preference data.

Approximately 95 structured governance scenarios are constructed across four tiers:

Tier	Count	Purpose	Design
1	∼55	Parameter isolation	Single feature varies; all others at baseline
2	∼28	Joint estimation + scale anchoring	Realistic multi-penalty combinations
3	∼34	Behavioural probes	Matched pairs testing cognitive biases
4	1	Historical calibration	Qantas Nov 2023 AGM (out-of-sample)

Table 4: Scenario battery design.

Each scenario is a complete game state: a decision node, a vote outcome, CEO status, review status, and a set of feasible actions. The LLM receives a natural-language description and returns Likert ratings (1–5) for each feasible action. Parameters are identified by **controlled variation**: pairs of scenarios that differ in exactly one game-state feature isolate the corresponding parameter's effect on the LLM's severity ratings.

For each scenario, 40 independent LLM calls are made (with randomised factor presentation order), producing $95 \times 40 = 3{,}800$ observations. This provides sufficient replication to estimate within-scenario variance and detect systematic biases.

The Tier 3 behavioural probes test for cognitive biases (loss aversion, self-assessment bias, the Ikea effect, optimism bias, and non-linearity around vote thresholds) that would

violate model assumptions. These do not enter the estimation; they serve as diagnostic checks on whether the LLM exhibits biases that require correction.

The single Tier 4 scenario (the actual Qantas November 2023 AGM at $V = 0.83$) provides an out-of-sample validation: does the model's predicted action ranking match what actually happened (the Board chose to commission a governance review)?

6.2.3 Bayesian Estimation via Ordinal Probit

The LLM produces ordinal Likert ratings (1–5), which are observations of an underlying latent utility discretised into five categories. The natural statistical model is the **ordinal probit**, which maps latent utility to observed categories via cumulative normal thresholds.

The latent utility for each (scenario, action) pair s is:

$$\mu_s = \boldsymbol{\phi}_s \cdot \mathbf{w} + \text{anchored}_s - \mathbf{1}[\text{strike}] \cdot w_{\text{v\_strike}} \cdot x_s^{\text{strike}} - \mathbf{1}[\text{overwhelming}] \cdot w_{\text{v\_over}} \cdot x_s^{\text{over}} \quad (4)$$

where $\boldsymbol{\phi}_s$ is the basis function row for pair s and $\mathbf{w}$ is the vector of 10 action-varying utility weights.

Raw μ values can span 20+ units (dominated by vote penalties), placing most observations in saturated probit tails with zero gradient. A pre-computed scale factor normalises $\eta = (\mu + \text{RE})/\mu_{\text{scale}}$ so the probit sees values in approximately $[-3, 3]$.

The observation model is:

$$y_n \sim \text{OrderedProbit}(\eta_{s(n)},\ \mathbf{c}) \quad (5)$$

where $\mathbf{c} = (c_1, c_2, c_3, c_4)$ are the ordered cutpoints and η includes a scenario-level random intercept (non-centred parameterisation: $\sigma_{\text{scenario}} \cdot z_{\text{scenario}},\ z \sim \mathcal{N}(0,1)$) to capture unmodelled scenario heterogeneity.

All 10 utility weights have **lognormal priors** centred at the expert-set defaults (Table 3), with prior standard deviation 1.0 on the log scale, giving approximately a 2.7× range per standard deviation on the ratio scale. The lognormal family naturally enforces the positivity constraint $w_k > 0$. An ordering constraint $w_{\text{removal}} > w_{\text{remove\_ceo\_overwhelming}}$ is enforced via a reparameterisation $w_5 = w_6 + \delta$, $\delta \sim \text{Lognormal}(0.26, 1.0)$.

The cutpoints are reparameterised for numerical stability:

$$c_1 = 3 \cdot \tanh(b_{\text{raw}}), \qquad c_{g+1} = c_g + 0.25 + 2.0 \cdot \text{logit}^{-1}(g_{\text{raw},g}) \quad (6)$$

bounding the base location to $[-3, 3]$ and gaps to $[0.25, 2.25]$, which avoids cutpoint degeneracy and numerical overflow during warmup.

The model is estimated using Stan MCMC with 4 chains × 2,000 sampling draws (plus 1,000 warmup) = 8,000 posterior draws, `adapt_delta` = 0.99, and `max_treedepth` = 15. Convergence is assessed via $\hat{R}$, bulk effective sample size (> 400 per parameter), and the number of divergent transitions.

The estimation produces:

- An 8,000 × 10 matrix of posterior weight draws $\mathbf{W} = \{w^{(i)}\}_{i=1}^{8000}$.
- Posterior summaries (mean, standard deviation, 95% credible intervals) for all 12 parameters.
- The full posterior covariance matrix for downstream uncertainty propagation.

6.2.4 From Posterior Draws to Board Decision Probabilities

The posterior weight draws feed directly into the game tree. At each terminal node, Board expected utility for a single posterior draw i is:

$$\mathrm{EU}_B^{(i)} = \mathbf{w}^{(i)} \cdot \boldsymbol{\phi} + \text{anchored} \tag{7}$$

Because each posterior draw produces different weight values, each draw produces a different EU, reflecting uncertainty in the Board's preferences.

EU arrays propagate backward from terminals through the tree:

- **Terminal nodes:** compute $\mathrm{EU}_B^{(i)}$ for each draw i.
- **Chance nodes** (vote V, review R): weighted sum over child EU arrays using per-draw probabilities: $\mathrm{EU}_{\text{parent}}^{(i)} = \sum_j p_j^{(i)} \cdot \mathrm{EU}_{\text{child}_j}^{(i)}$.
- **Non-Board decision nodes** (ASA, CEO): weighted sum using action probabilities from ARA predictive distributions with per-draw Dirichlet epistemic noise.
- **Board decision nodes:** argmax-count (see below).

6.2.5 Argmax-Count: How Board Probabilities Emerge

At each Board decision node, the probability of each action is determined by **per-draw utility maximisation**. Let $\mathrm{EU}_a^{(i)}$ denote the expected utility of action a under posterior draw i. The Board's decision probability is:

$$P(a) = \frac{\left|\{i : a \in \arg\max_{a'} \mathrm{EU}_{a'}^{(i)}\}\right| + \alpha}{N + K\alpha} \tag{8}$$

where $N = 8{,}000$ is the number of posterior draws, K is the number of feasible actions, and $\alpha = 1.0$ is a Laplacian smoothing constant ensuring no action receives exactly zero probability.

Interpretation: On posterior draw i, the Board has specific weight values $\mathbf{w}^{(i)}$. Given those weights, one action has the highest EU, and that action gets the Board's "vote" for draw i. The final probability is the fraction of draws on which each action wins.

This is fundamentally different from a softmax decision rule $P(a) \propto \exp(\lambda \cdot \mathrm{EU}(a))$, which requires choosing a rationality parameter λ whose correct value is unknown. The distinction matters both practically and theoretically. A softmax rule treats decision noise as *aleatory*: the decision-maker is assumed to randomise over actions with probabilities that increase in expected utility, governed by a "rationality temperature" λ. Low λ produces near-uniform randomisation (bounded rationality); high λ approaches deterministic maximisation. The problem is that λ is not identified from the data: different values produce materially different action probabilities, and there is no principled way to calibrate it without assuming the very behavioural model one is trying to estimate.

Argmax-count sidesteps this entirely. It treats the uncertainty as *epistemic* rather than aleatory: the Board is assumed to maximise expected utility given its true preferences, but the analyst is uncertain about what those preferences are. Each posterior draw $\mathbf{w}^{(i)}$ represents a plausible preference vector; on each draw, the Board acts deterministically (choosing the action with highest EU), and the analyst's uncertainty about preferences generates a *distribution* over which action is chosen. When the posterior is tight

(parameters well-identified), the probabilities concentrate on one action. When the posterior is diffuse, they spread across actions. The data determines the degree of decisiveness, not an ad hoc temperature parameter.

This approach is consistent with the ARA framework as developed by Rios Insua, Banks, and Rios (2009), where uncertainty about an opponent's action arises from the analyst's uncertainty about the opponent's utility function and beliefs, not from assumed irrationality on the opponent's part. The ARA textbook treats each opponent as a subjective expected utility maximiser whose parameters are unknown to the analyst; the analyst's predictive distribution over the opponent's action is then obtained by integrating over the posterior on those parameters, exactly the operation that argmax-count implements via Monte Carlo. The result is a decision rule that is fully Bayesian, requires no tuning parameters beyond the posterior itself, and degrades gracefully as parameter uncertainty increases.

After computing the mean action probabilities from argmax-count, per-draw Dirichlet noise is added to capture small-sample epistemic uncertainty:

$$\boldsymbol{\alpha}_{\text{Dir}} = \frac{\mathbf{p}}{\sum p_a} \cdot C, \qquad \mathbf{p}^{(i)}_{\text{Dir}} \sim \text{Dirichlet}(\boldsymbol{\alpha}_{\text{Dir}}) \tag{9}$$

where $C = 20.0$ is the concentration sum. These per-draw Dirichlet probabilities propagate all actor EU streams (Board, ASA, CEO) through the node.

6.2.6 Board-Focal Tree Recursion

The full backward recursion for the Board-focal tree (in the "CEO stayed" scenario) is:

$$U_B^{(D_1)}(h_0) = \max_{d_1 \in \mathcal{D}_1} U_B^{(A_2)}(h_0 \cup \{D_1 = d_1\}) \tag{10}$$

$$U_B^{(A_2)}(h) = \sum_{a \in \mathcal{A}_2} p_B(A_2 = a \mid h) \cdot U_B^{(V)}(h \cup \{A_2 = a\}) \tag{11}$$

$$U_B^{(V)}(h) = \frac{1}{M_V} \sum_{m=1}^{M_V} U_B^{(D_4)}(h \cup \{V_m\}) \tag{12}$$

At the CEO's decision node (opponent: predictive distribution or skip if absent):

$$U_B^{(D_4)}(h) = \begin{cases} \displaystyle\sum_{d_4} p_B(D_4 = d_4 \mid h) \cdot U_B^{(D_{\text{rev}})}(h \cup \{D_4 = d_4\}) & \text{if CEO present} \\ U_B^{(D_{\text{rev}})}(h) & \text{if CEO absent} \end{cases} \tag{13}$$

At the Board's post-AGM response (focal: maximise):

$$U_B^{(D_{\text{rev}})}(h) = \max_{d \in \mathcal{D}_{\text{rev}}(S)} U_B^{(R)}(h \cup \{D_{\text{rev}} = d\}) \tag{14}$$

Review findings are integrated via Monte Carlo:

$$U_B^{(R)}(h) = \frac{1}{M_R} \sum_{m=1}^{M_R} U_B^{(\text{post-review})}(h \cup \{R_m, C_{\text{direct}}\}) \tag{15}$$

with a conditional post-review round if the review finding is adverse and the CEO is still present:

$$U_B^{(\text{post-review})}(h) = \begin{cases} U_B^{(D_4')}(h) & \text{if review adverse } \wedge \text{ CEO present} \\ u_B(Z(h,S)) & \text{otherwise} \end{cases} \tag{16}$$

The Board's optimal initial action is then:

$$d_1^* \in \arg\max_{d_1 \in \mathcal{D}_1} \frac{1}{N} \sum_{i=1}^{N} U_B^{(A_2)}(h_0 \cup \{D_1 = d_1\};\, \theta^{(i)}) \tag{17}$$

6.2.7 ARA Predictive Distributions for Opponents

At each opponent decision node X owned by player $j \in \{\text{ASA}, \text{CEO}\}$, the Board forms a predictive distribution over the opponent's action:

$$p_B(X = x \mid h) = \frac{1}{K} \sum_{k=1}^{K} \mathbf{1}\left[x \in \arg\max_{x' \in \mathcal{X}(h)} \Psi_j(x';\, h,\, \Theta_j^{(k)})\right] \tag{18}$$

where $K = 200$ opponent parameter samples $\Theta_j^{(k)} \sim p_B(\Theta_j)$ are drawn from the Board's prior over the opponent's preferences, and $\Psi_j(x;\, h,\, \Theta_j) = \frac{1}{R} \sum_{r=1}^{R} u_j(Z^{(r)};\, \Theta_j)$ with $R = 20$ stochastic rollouts per action. Rollouts use fixed policies for non-evaluated actors and the Board's (possibly biased) model of Nature.

6.2.8 Stochastic Decision Modelling

The stochasticity in the Board's modelled decisions arises from three distinct sources:

- **Parameter uncertainty**: the posterior distribution over utility weights $\mathbf{w}$ reflects genuine uncertainty about how the Board trades off competing concerns.
- **Opponent uncertainty**: the decisions of other stakeholders (CEO, ASA, shareholders) are themselves uncertain and modelled via ARA predictive distributions.
- **Nature uncertainty**: chance outcomes (vote fraction, review findings, CAR) are drawn from calibrated probability models integrated via Monte Carlo.

The total computational budget for a single solve is $|\mathcal{D}_1| \times N$ tree evaluations, with $K \times |\mathcal{A}| \times R$ rollouts per predictive distribution call. Default values are $N = 500$ belief draws, $K = 200$ opponent samples, $R = 20$ rollouts, $M_V = 50$ vote samples, and $M_R = 20$ review samples.

6.3 Modelling the Australian Shareholders Association

The ASA has a simpler decision space: their only decision is **whether to recommend voting against the remuneration report** at each of five game-tree nodes (corresponding to the five possible game states at the point the ASA must decide). Because the decision space is binary and the historical strike rate is very high, the modelling approach differs from the Board: we start with probabilities, then reverse-engineer a utility function that is consistent with those probabilities.

6.3.1 The Coherence Problem

An ARA game tree requires two kinds of numerical input for each non-focal actor: **utility parameters** (weights that determine how the actor trades off competing concerns) and **action probabilities** (the distribution over feasible actions at each decision node). These two inputs are not independent. If ASA's utility assigns weight w to board accountability and weight v to mobilisation cost, then the probability of a strike recommendation is *determined* by those weights applied to the game state. Assigning probabilities and utility weights separately risks **incoherence**: the probabilities may be inconsistent with the utility function, or vice versa.

The pipeline enforces coherence by deriving action probabilities *from* the utility function via simulation. The final $P(\text{strike})$ at each A_2 node is the fraction of stochastic weight draws for which $\text{EU}(\text{strike}) > \text{EU}(\text{no\_strike})$.

6.3.2 The ASA Utility Function

The ASA utility function is specified in two stages. The **pre-estimation scaffold** is a seven-dimensional weighted assessment model (dimensions FW, PPL, TD, EGR, BA, OL, PF, each scored on a $[1, 5]$ Likert scale with charter-derived weights summing to 1.0) that is used only to generate the elicitation scenarios and define the structural phi-function basis over game-state features. Once the digital-twin Likert responses are collected, the model is re-parameterised in terms of **eight context and interaction weights** that are directly identified by the scenario design, and these weights are estimated via the same Stan ordinal-probit procedure used for the Board. Posterior means and 95% credible intervals are shown in Table 5; these are the coefficients that actually enter the ARA engine.

Parameter	Description	Post. mean	95% CI
$w_{\text{ctx\_inact}}$	Penalty for visible Board inaction (no review, no sack)	2.54	[1.96, 3.08]
$w_{\text{ctx\_dep}}$	Credit for CEO having departed (voluntary or forced)	0.60	[0.20, 1.07]
$w_{\text{ctx\_rev}}$	Credit for Board commissioning a governance review	0.30	[0.07, 0.70]
$w_{\text{strike\_cost}}$	Net mobilisation cost of recommending a strike	0.14	[0.04, 0.28]
$w_{\text{strike\_vs\_pass}}$	Marginal value of striking when Board has not acted	7.02	[6.52, 7.57]
$w_{\text{dep\_dampens}}$	CEO departure reduces marginal value of striking	0.80	[0.62, 0.96]
$w_{\text{sack\_dampens}}$	Board-forced CEO exit further reduces strike value	0.02	[0.008, 0.051]
w_{cred}	Repeat-game credibility value of striking in a high-profile case	1.10	[0.93, 1.27]

Table 5: ASA utility function: posterior means and 95% credible intervals for the reparameterised context and interaction weights, estimated via Stan ordinal probit from the digital-twin elicitation data. The dominant term is $w_{\text{strike\_vs\_pass}}$, consistent with the empirical base rate that ASA almost always recommends a strike when the Board has taken no visible action.

The utility function is:

$$u_A(Z) = \sum_{d \in \mathcal{D}} w_d \cdot \text{clip}\big(s_d^{\text{base}}(Z) + \Delta s_d(Z),\ 1,\ 5\big) - w_{\text{mob}} \cdot \mathbf{1}[a_2 = \text{rec\_strike}] \tag{19}$$

where:

- $s_d^{\text{base}}(Z)$ are base Likert scores per dimension, determined by the game state at A_2 (CEO status and Board action d_1).
- $\Delta s_d(Z)$ are post-A_2 adjustments triggered by downstream outcomes: a strike adds $\Delta\text{BA} = +1.5$, $\Delta\text{OL} = +1.0$; an overwhelming vote adds $\Delta\text{BA} = +1.0$, $\Delta\text{OL} = +0.5$; involuntary CEO removal adds $\Delta\text{BA} = +1.0$, $\Delta\text{FW} = +0.5$; a negative review adds $\Delta\text{TD} = +1.0$, $\Delta\text{EGR} = +0.5$; and market alignment (ASA recommended strike and strike occurred) adds $\Delta\text{OL} = +1.0$, $\Delta\text{PF} = +0.5$.
- $w_{\text{mob}} = 0.3$ is the mobilisation cost of recommending a strike.

6.3.3 Historical Base Rates

The starting point for ASA calibration is observed behaviour. ASA's track record in 15 comparable headline governance incidents in Australian listed companies provides an empirical anchor:

Board response	**Incidents**	**ASA rec. strike**	**Rate**
Board did nothing	10	9	90.0%
Board commissioned review or CEO resigned	3	3	100.0%
Board sacked CEO	2	2	100.0%
Overall	**15**	**14**	**93.3%**

Table 6: ASA historical strike recommendation rates in headline governance crises (aggregated from the 14 eligible observations in Table 7).

The underlying case-level data is presented in Table 7. Qantas itself is excluded from these counts because the modelled entity cannot inform its own prior.

These base rates establish that ASA almost always recommends a strike in headline governance crises. However, they are insufficient for three reasons: (1) with only 15 observations, the conditional rates are too noisy to distinguish between nodes (a binomial 95% confidence interval for 2/2 extends from 34% to 100%); (2) the base rates do not identify the utility weights, since infinitely many weight vectors produce the same 93% aggregate rate; and (3) the Qantas crisis is among the most severe in Australian corporate history, so the average base rate likely understates the strike probability for this specific case.

6.3.4 Scaffolded Elicitation

Because direct estimation from historical data is infeasible, the pipeline uses a **scaffolding approach**: multiple layers of structured elicitation that progressively constrain the parameter space until the utility weights are identified.

Company	Year	Board Action	ASA Against	Crisis Type
Rio Tinto	2021	D3 (CEO exit)	Yes	Juukan Gorge cultural heritage destruction
AGL Energy	2021	D1 (review)	Yes	Climate/emissions governance
Westpac	2021	D1 (review)	Yes	AUSTRAC money-laundering scandal
IAG	2021	D1 (review)	Yes	Insurance pricing misconduct
Downer EDI	2022	D0 (no action)	Yes	Governance concerns
ASX Ltd	2022	D3 (CEO exit)	Yes	Technology governance failure (CHESS)
Harvey Norman	2023	D0 (no action)	Yes	Wage theft scandal
Fortescue	2023	D0 (no action)	No	Headline incident; only non-recommendation
Woolworths	2023	D0 (no action)	Yes	Cost-of-living pricing scandal
Elders	2023	D0 (no action)	Yes	Governance concerns
Mineral Resources	2023	D0 (no action)	Yes	Governance concerns
Perpetual	2023	D0 (no action)	Yes	Governance concerns
Atlas Arteria	2023	D0 (no action)	Yes	Governance concerns
Macquarie Group	2023	D0 (no action)	Yes	Governance concerns

Table 7: ASA headline governance incidents (n = 14, excluding Qantas). Board action codes: D0 = no action, D1 = review/governance reform, D3 = CEO transition. Fortescue is the only headline-incident case where ASA did not recommend against.

Step 1: Range elicitation. An LLM (gpt-4o-mini), prompted as a calibrated ASA company monitor, provides three probability values for $P(\text{strike})$ at each of the five A_2 nodes: LOW (lowest plausible), BEST (point estimate), and HIGH (highest plausible). The prompt anchors the LLM to the historical base rate (93.3%) and to case-specific facts. It also elicits floor probabilities: an expected floor (the minimum $P(\text{strike})$ across all scenarios, typically 0.88–0.95) and an absolute floor (hard lower bound, typically 0.80–0.88). This is repeated with 30 independent LLM draws to capture the LLM's own uncertainty.

Step 2: Gap elicitation. Rather than absolute probabilities (which are hard to calibrate in the high-probability regime where all values cluster near 95%), the LLM assesses **pairwise gaps**: how much each additional Board action reduces $P(\text{strike})$:

Gap	Comparison	What it measures
Departure	Node 3 vs Node 1	CEO departure reduces $P(\text{strike})$?
Review (CEO stays)	Node 3 vs Node 4	Review reduces $P(\text{strike})$ when CEO stays?
Review (CEO resigned)	Node 1 vs Node 2	Review further reduces $P(\text{strike})$ after departure?
Sacking signal	Node 2 vs Node 5	Board-forced exit vs voluntary departure + review?

Table 8: Pairwise gap elicitations for ASA strike probabilities.

The prompt emphasises that the remuneration vote is **retrospective**: Board actions taken after the FY23 pay period are forward-looking signals that may slightly moderate ASA's position, but do not change the historical pay structure being voted on. This is why gaps are small (typically 1–8 percentage points).

Step 3: Constrained optimisation. The range and gap elicitations provide overlapping, potentially inconsistent constraints. These are reconciled via constrained optimisation (SLSQP), minimising squared distance from elicited best estimates subject to:

1. **Range constraints:** each target probability lies within its elicited [low, high] range.
2. **Floor constraint:** all targets exceed the expected floor probability.
3. **Monotonicity:** the ranking respects the a priori ordering:

$$P_3 > P_1 \geq P_4 > P_2 > P_5 \tag{20}$$

   where Node 3 (CEO stays, Board does nothing) produces the highest $P(\text{strike})$ and Node 5 (Board sacked CEO) produces the lowest. A minimum gap of 0.5 percentage points is enforced between consecutive ranked nodes. The ordering $P_1 \geq P_4$ (i.e. that a Board-initiated review provides more mitigation when the CEO has already resigned than when the CEO remains) encodes a substantive governance claim: if the CEO resigns, the Board is perceived as having been passive and in thrall of a dominant CEO, so commissioning an independent review is the strongest available signal that the Board is meeting its duties and taking ownership of governance failures. By contrast, if the CEO has been fired, the Board has already exceeded the passivity threshold through decisive action, and commissioning a review is framed by shareholders and advisers as supplementary rather than necessary to demonstrate minimum accountability. The same logic explains why P_5 (Board sacked CEO) is the lowest: the most dramatic Board action leaves the least residual governance deficit for the vote to punish.
4. **Gap constraints:** pairwise gaps lie within the elicited [low, high] ranges.

This produces a monotone target probability ladder, e.g.:

Node 3 (stay + nothing):	0.980
Node 1 (resign + nothing):	0.967
Node 4 (stay + review):	0.962
Node 2 (resign + review):	0.957
Node 5 (sacked):	0.927

6.3.5 Random Utility Model: From Utility to Probability

The pipeline uses a **random utility model** (McFadden, 1974) to connect utility weights to action probabilities. Utility weights are not point values but random variables drawn from truncated normal distributions. For each simulation draw, the weights are sampled, expected utilities are computed for both actions, and the action with higher EU is selected. The probability of a strike recommendation is the fraction of draws where EU(strike) > EU(no_strike).

Formally, let $w_k \sim \text{TruncNormal}(\mu_k, \sigma_k, 0.5, 10.0)$ for each interaction parameter k. At each A_2 node, the delta-EU is:

$$\Delta\text{EU} = \sum_k w_k \cdot \Delta\phi_k \tag{21}$$

where $\Delta\phi_k = \phi_k(\text{rec\_strike}) - \phi_k(\text{no\_strike})$ is the basis function difference. Strike wins when $\Delta\text{EU} > 0$.

The utility parameters decompose into two classes:

- **Context parameters** ($w_{\text{ctx\_inaction}}$, $w_{\text{ctx\_departure}}$, $w_{\text{ctx\_review}}$) fire equally for both actions, cancel in ΔEU, and have no effect on action probabilities.
- **Interaction parameters** (5 parameters) fire only for rec_strike and drive the decision:

Parameter	$\Delta\phi$	Interpretation
$w_{\text{strike\_cost}}$	-1 (all nodes)	Net mobilisation cost of striking
$w_{\text{strike\_vs\_passive}}$	$+1$ (Board inactive)	Value of striking against a passive Board
$w_{\text{departure\_dampens}}$	-1 (CEO departed)	CEO departure reduces strike value
$w_{\text{sack\_dampens}}$	-1 (CEO sacked)	Board-forced exit further reduces strike value
$w_{\text{credibility\_signal}}$	$+1$ (all nodes)	Repeat-game credibility value of striking

Table 9: ASA interaction parameters and their delta-phi values.

6.3.6 Two-Stage Optimisation

The pipeline optimises the 10 parameters (5 means + 5 standard deviations of the truncated normal distributions) to match the target probabilities from Step 3.

Stage 1: Analytical CLT (fast initial guess). Since ΔEU is a sum of independent truncated normals, its distribution is approximately normal by the Central Limit Theorem:

$$P(\text{strike}) \approx \Phi\left(\frac{\mathbb{E}[\Delta\text{EU}]}{\sqrt{\text{Var}[\Delta\text{EU}]}}\right) \tag{22}$$

where Φ is the standard normal CDF. This gives a smooth, differentiable loss function. Multi-start L-BFGS-B (50 random restarts) finds a good initial guess.

Stage 2: Monte Carlo refinement with common random numbers. The CLT approximation is systematically biased when parameters are at truncation bounds (e.g. TruncNormal$(0.5, 2.1, 0.5, 10)$ is heavily right-skewed). Stage 2 pre-draws 100,000 uniform random numbers (common random numbers, fixed seed), then transforms them via the inverse CDF of the truncated normal. This makes the Monte Carlo loss function **deterministic** (same draws every call) and **unbiased** (exact truncated normal shape). Nelder-Mead refines from the Stage 1 solution. Typical results: Stage 1 loss ~ 0.00013, Stage 2 loss ~ 0.00008; maximum validation error ~ 0.8 percentage points.

6.3.7 Beta Distributions for Engine Integration

The ARA engine models ASA stochastically when ASA is not the focal actor. The Beta distribution is the natural choice: it is the conjugate prior for a Bernoulli outcome (strike/no strike), is bounded on $[0, 1]$, and is fully characterised by two parameters.

Node-specific Beta distributions are derived using method of moments:

$$n_{\text{eff}} = \frac{\hat{p}(1-\hat{p})}{\hat{\sigma}^2} - 1, \qquad \alpha = \hat{p} \cdot n_{\text{eff}}, \qquad \beta = (1-\hat{p}) \cdot n_{\text{eff}} \tag{23}$$

Node	P(strike)	Elicited SD	n_{eff}	Beta	95% CI
Node 3 (stay + nothing)	0.979	0.009	235	Beta(230, 5)	[0.957, 0.993]
Node 1 (resign + nothing)	0.966	0.014	175	Beta(169, 6)	[0.934, 0.987]
Node 4 (stay + review)	0.965	0.015	143	Beta(138, 5)	[0.929, 0.989]
Node 2 (resign + review)	0.946	0.015	220	Beta(208, 12)	[0.912, 0.971]
Node 5 (sacked)	0.924	0.028	92	Beta(85, 7)	[0.862, 0.969]

Table 10: Calibrated Beta distributions for ASA strike probabilities at each A_2 node.

where $\hat{p}$ is the MC-optimised P(strike) and $\hat{\sigma}$ is the standard deviation from the LLM elicitation draws (capturing genuine epistemic uncertainty). The effective sample size n_{eff} is clipped to $[20, 500]$ to prevent degenerate distributions.

The sacking scenario (Node 5) has the widest uncertainty (SD = 0.028, $n_{\text{eff}} = 92$) because there are only 2 historical observations of Board-forced exits. The Board-inaction scenario (Node 3) has the narrowest uncertainty (SD = 0.009, $n_{\text{eff}} = 235$) because this is consistently rated as near-certain strike territory.

In the engine, when ASA is not the focal actor, P(strike) is sampled per rollout:

$$p_{\text{strike}} \sim \text{Beta}(\alpha_d, \beta_d), \qquad A_2 = \begin{cases} \text{rec\_strike} & \text{w.p. } p_{\text{strike}} \\ \text{no\_strike} & \text{w.p. } 1 - p_{\text{strike}} \end{cases} \tag{24}$$

Each rollout draws a different p_{strike}, so the engine naturally represents uncertainty about ASA's behaviour. When ASA *is* the focal actor, the full seven-dimensional utility function (Equation 19) is evaluated through the tree recursion, and ASA maximises:

$$a_2^*(d_1) \in \arg\max_{a \in \mathcal{A}_2} \frac{1}{N} \sum_{i=1}^{N} U_A^{(V)}(h \cup \{A_2 = a\};\, \theta^{(i)}) \tag{25}$$

6.4 Modelling Shareholders

Unlike the Board, ASA, and CEO, who are modelled as strategic decision-makers with utility functions, shareholders are modelled as **Nature** in ARA terminology. This means we do not attribute a utility function or strategic intent to the aggregate shareholder body. Instead, the vote outcome is treated as a **stochastic chance node**: a conditional probability distribution over the vote fraction, driven by upstream decisions and latent belief states. This is appropriate because the vote aggregates millions of individual decisions by heterogeneous shareholders with diverse information sets, and no single utility function can represent the collective.

6.4.1 Historical Qantas AGM Voting Data

The empirical foundation for the shareholder vote model is the time series of Qantas AGM remuneration votes, which anchors the Stan state-space belief model and calibrates the vote escalation pattern informing the B_{mkt} posterior draws.

The pre-crisis votes (8–10% against) establish the baseline opposition level that anchors the Stan belief model's intercept α_{rem}, while the 2023 crisis vote provides the out-of-sample observation incorporated via sequential Monte Carlo at checkpoint C_3.

Year	Vote Against Rem. (%)	Vote Against Chair (%)	Context
2020	8.93	n/a	Post-COVID; moderate dissent
2021	9.99	n/a	Stable; no headline incident
2022	9.38	2.04	Stable; pre-crisis
2023	**82.93**	n/a	Crisis AGM: ACCC ghost flights, Senate inquiry, CEO resignation

Table 11: Qantas AGM remuneration vote history, 2020–2023. The 73.6 percentage point escalation from 2022 to 2023 is the largest year-on-year increase in the panel and one of the largest in ASX history. The 82.93% vote exceeded both the 25% first-strike and 50% "overwhelming" thresholds.

6.4.2 The Logit-Normal Vote Model

The vote fraction $V \in (0, 1)$ follows a logit-normal distribution:

$$\text{logit}(V) \sim \mathcal{N}(\mu_V,\ \sigma_V^{(i)}), \qquad V = \text{expit}(\text{logit}(V)) = \frac{1}{1 + e^{-\text{logit}(V)}} \tag{26}$$

The location parameter μ_V is composed of additive terms:

$$\mu_V = \alpha_V^{(i)} + B_{\text{agm}}^{(i)} \tag{27}$$

where:

$$B_{\text{agm}} = B_{\text{mkt}}^{(i)} + \gamma_A^{(i)} \cdot \mathbf{1}[A_2 = \text{rec\_strike}] + \gamma_{AH}^{(i)} \cdot \mathbf{1}[A_2 = \text{rec\_strike}] \cdot \mathbf{1}[\text{headline}] + \gamma_D^{(i)} \cdot f(D_1) \tag{28}$$

The superscript (i) denotes posterior draw i from the upstream Bayesian estimation. The parameters are:

Symbol	Name	Source	Description
$\alpha_V^{(i)}$	Baseline intercept	Stan posterior	Baseline opposition on the logit scale
$B_{\text{mkt}}^{(i)}$	Market belief	Stan posterior	Latent shareholder distrust at checkpoint date
$\gamma_A^{(i)}$	ASA effect (baseline)	First-diff. OLS + prior	Logit shift from ASA mobilisation (non-crisis)
$\gamma_{AH}^{(i)}$	ASA effect (headline)	Subgroup OLS + prior	Additional logit shift in headline-incident cases
$\gamma_D^{(i)}$	Governance sensitivity	Stan posterior	Scaling coefficient for governance action effect
$\sigma_V^{(i)}$	Vote noise	Stan posterior	Logit-scale aleatoric uncertainty
$f(D_1)$	Governance effect	Sampled at runtime	Effectiveness of Board's governance reform

Table 12: Vote model parameter definitions.

The logit-normal is the natural choice for a vote fraction: it is bounded on $(0, 1)$ by construction, admits additive effects on the logit scale (consistent with the upstream Stan state-space model that anchors $\lambda_{\text{rem}} = 1$), and has flexible shape through its regression-style predictors.

6.4.3 Parameter Estimation Pipeline

The vote model parameters flow through a three-stage estimation pipeline:

Stage 1: Upstream Stan models. Two Bayesian state-space models produce the foundational estimates. A **media measurement model** separates latent media coverage from intensity using sparse monthly observations via an AR(1) process:

$$\log M_t = \mu_{\log M} + \phi(\log M_{t-1} - \mu_{\log M}) + \sigma_{\log M} \cdot z_t \tag{29}$$

The log-differences $\Delta \log M_t$ become the media shock series that enters the **belief dynamics model**, which tracks a latent shareholder distrust state B_t via AR(1) dynamics:

$$B_t = \rho \cdot B_{t-1} + \beta \cdot \text{shock}_t + \sigma_B \cdot z_t \tag{30}$$

The belief state is identified through three observation channels: abnormal returns ($\text{abret}_t \sim \mathcal{N}(\lambda_r \cdot B_t, 1)$), remuneration vote ($\text{logit}(y_{\text{rem}}) \sim \mathcal{N}(\alpha_{\text{rem}} + B_t, \sigma_{\text{rem}})$ with $\lambda_{\text{rem}} = 1$ fixed for scale anchoring), and chair vote ($\text{logit}(y_{\text{chair}}) \sim \mathcal{N}(\alpha_{\text{chair}} + \lambda_{\text{chair}} \cdot B_t, \sigma_{\text{chair}})$). Fixing $\lambda_{\text{rem}} = 1$ pins the belief state to the logit scale of remuneration opposition, so the game tree operates on exactly the same scale as the upstream estimation.

Stage 2: ASA mobilisation effects. The effects γ_A (baseline) and γ_{AH} (headline interaction) are estimated from a cross-company panel of 36 voting recommendations using a first-differences specification:

$$\Delta\text{logit}(y_{\text{rem}}) = b_0 + b_1 \cdot \Delta\text{asa\_against} + b_2 \cdot \Delta\text{asa\_against} \cdot \text{headline}_t + b_3 \cdot \Delta \log(\text{mkt\_cap}) + \varepsilon \tag{31}$$

First-differencing eliminates time-invariant company characteristics (governance quality, ownership structure, investor base) that confound the ASA effect in a levels regression. The OLS estimates b_1 and b_2 are converted to truncated Normal priors:

$$\gamma_A \sim \mathcal{N}^+(b_1, \sqrt{\text{SE}_1^2 + \tau_A^2}), \qquad \gamma_{AH} \sim \mathcal{N}^+(b_2, \sqrt{\text{SE}_2^2 + \tau_{AH}^2}) \tag{32}$$

where τ terms are cross-company heterogeneity adjustments and the truncation at zero encodes the domain constraints that ASA mobilisation cannot reduce opposition.

Stage 3: Checkpoint construction. Posterior draws are combined with shock priors to produce belief checkpoints at five critical dates:

Checkpoint	Date	Context	Key information
C_{pre}	Pre-crisis	Baseline	Common beliefs
C_0	2023-10-01	Post-CEO resignation	Management has private ASA signal
C_1	2023-10-10	Review announced	Common knowledge of review
C_2	2023-10-18	ASA mobilisation public	$\gamma_A + \gamma_{AH}$ become common knowledge
C_3	2023-11-03	Post-AGM (82.9% observed)	Posterior sharpened via SMC

Table 13: Belief checkpoint timeline. Each checkpoint contains 500 draws of all vote model parameters.

At C_3, the observed 82.9% vote is incorporated via sequential Monte Carlo (SMC) with tempered likelihoods to maintain particle diversity, replacing simple importance resampling which risks effective sample size collapse when the observed vote lies in the prior's extreme tail.

6.4.4 Governance Effect

The governance effect $f(D_1)$ is drawn from a Uniform distribution whose bounds depend on the Board's action:

$$f(D_1) = \begin{cases} 0 & D_1 = \text{D0\_minimal} \\ U(0,1) & D_1 = \text{D1\_review} \\ U(-1,0.5) & D_1 = \text{D3\_ceo\_transition} \end{cases} \tag{33}$$

When multiplied by γ_D (negative from the posterior), a review always reduces protest (positive f times negative γ_D lowers B_{agm}), while CEO exit has an ambiguous effect: the asymmetric bounds $U(-1, 0.5)$ encode a prior that amplification is roughly three times as likely as mitigation, but do not foreclose the possibility that a well-managed transition partially mollifies shareholders.

This encoding is calibrated from a cross-company panel of 36 ASX remuneration votes (Table 14), which also underpins the ASA mobilisation estimation and structural floor calibration.

Company	Code	Year	Rem. Against (%)	Strike	ASA	Proxy	Headline	Board	Δvote (pp)
Rio Tinto	RIO	2021	61.0	Y	Y	Y	Y	D1	+49.0
AGL Energy	AGL	2021	31.0	Y	Y	Y	Y	D0	+16.0
QBE Insurance	QBE	2021	44.0	Y	Y	Y	N	D1	+36.0
Scentre Group	SCG	2021	51.0	Y	Y	Y	N	D1	+41.0
Westpac	WBC	2021	30.0	Y	Y	Y	Y	D1	+12.0
IAG	IAG	2021	57.3	Y	Y	Y	Y	D1	+35.3
Dexus	DXS	2021	66.0	Y	Y	Y	N	D0	+61.0
Link Admin	LNK	2021	63.0	Y	Y	Y	N	D0	+49.0
Platinum AM	PTM	2021	50.0	Y	Y	Y	N	D0	+38.0
Argo Inv.	ARG	2021	17.0	N	Y	N	N	D0	0.0
Downer EDI	DOW	2022	55.8	Y	Y	Y	Y	D0	+43.8
Blackmores	BKL	2022	43.4	Y	Y	Y	N	D0	+28.4
Newcrest	NCM	2022	37.0	Y	Y	Y	N	D0	+28.0
ASX Ltd	ASX	2022	31.0	Y	Y	Y	Y	D1	+26.0
Goodman Grp	GMG	2022	28.9	Y	Y	Y	N	D0	+7.9
Santos	STO	2022	25.3	Y	N	Y	N	D1	+17.3
Corporate Travel	CTD	2022	33.0	Y	Y	Y	N	D0	+22.0
BHP Group	BHP	2022	3.0	N	N	N	N	D0	+1.0
Rio Tinto	RIO	2022	15.7	N	N	N	N	D0	−45.3
Westpac	WBC	2022	12.0	N	N	N	N	D1	−18.0
Qantas	**QAN**	**2023**	**82.9**	**Y**	**Y**	**Y**	**Y**	**D1**	**+75.9**
Harvey Norman	HVN	2023	81.8	Y	Y	Y	Y	D0	+63.8
Fortescue	FMG	2023	52.0	Y	N	Y	Y	D0	+37.0
Treasury Wine	TWE	2023	46.0	Y	Y	Y	N	D0	+34.0
Tabcorp	TAH	2023	34.0	Y	Y	Y	N	D0	+25.0
Woolworths	WOW	2023	28.0	Y	Y	Y	Y	D0	+23.5
Elders	ELD	2023	62.7	Y	Y	Y	Y	D0	+48.7
Mineral Res.	MIN	2023	74.6	Y	Y	Y	Y	D0	+56.6
Computershare	CPU	2023	27.5	Y	Y	Y	N	D0	+23.5
Perpetual	PPT	2023	88.1	Y	Y	Y	Y	D0	+69.2
Healius	HLS	2023	55.0	Y	Y	Y	N	D0	+43.0
Sandfire Res.	SFR	2023	56.1	Y	Y	Y	N	D0	+41.1
BrainChip	BRN	2023	52.0	Y	Y	Y	N	D0	+34.0
Atlas Arteria	ALX	2023	51.0	Y	Y	Y	Y	D0	+37.0
Macquarie Grp	MQG	2023	25.4	Y	Y	Y	Y	D0	+14.4
NRW Holdings	NWH	2023	49.8	Y	Y	Y	Y	D0	+18.8

Table 14: Cross-company ASX remuneration voting panel ($n = 36$), 2021–2023. Columns: ASA = ASA recommended against; Proxy = institutional proxy adviser recommended against; Headline = headline governance incident; Board = Board corrective action (D0 = no action, D1 = review/reform, D3 = CEO transition); Δvote = year-on-year change in remuneration opposition. Qantas (bold) is the modelled entity.

Table 15 summarises the governance effects by Board action category:

Board action	n	**Mean Δvote**	**Interpretation**
D0 (no action)	25	+32.8 pp	Baseline escalation
D1 (review/announcement)	9	+14.9 pp	Most effective mitigation
D3 (CEO exit)	2	+46.7 pp	Ambiguous: confounded with crisis severity

Table 15: Governance effects on voting dissent, aggregated from the cross-company panel (Table 14).

The D1 effect (deficit of -17.9 pp relative to D0) is the most credible causal estimate because reviews are deployed across a range of crisis severities. The D3 excess (+13.9 pp over D0) is unreliable: the 4 CEO-exit cases are the most severe crises in the sample, and with $n = 4$ the selection bias cannot be separated from the causal effect.

The governance effect is drawn **once per belief draw** (epistemic uncertainty: "how effective is this particular reform?") and held fixed across the $M_V = 50$ vote samples (aleatoric uncertainty: "given the true effectiveness, what fraction of shareholders vote against?").

6.4.5 Structural Crisis Floor

In headline-incident cases, the model applies a structural floor:

$$V_{\text{floor}} \sim \text{Beta}(50, 150), \qquad V_{\text{final}} = \max(V_{\text{logit}},\ V_{\text{floor}}) \tag{34}$$

The floor has $\mathbb{E}[V_{\text{floor}}] = 0.25$ and 90% of its mass between 0.22 and 0.28. This encodes the domain knowledge that in the 36-observation panel, $P(\text{first strike} \mid \text{headline incident}) = 15/15 = 100\%$; the lowest headline-incident vote was MQG 2023 at 25.4%. A pure logit-normal with realistic parameter values assigns 3–14% probability to $V < 0.25$ even when the median vote is 50–60%, because the logit-normal has infinite support. The floor catches these implausible lower-tail draws. It is stochastic rather than fixed to avoid a mass point at exactly 0.25.

For non-crisis scenarios, no floor is applied and the logit-normal operates without constraint.

6.4.6 Derived Indicators and Thresholds

Two binary thresholds are derived from V_{final}:

$$\text{strike} = \mathbf{1}[V_{\text{final}} \geq 0.25], \qquad \text{overwhelming} = \mathbf{1}[V_{\text{final}} \geq 0.50] \tag{35}$$

A first strike (25%) triggers the "two strikes" rule under the Corporations Act 2001 §250U–250V. An overwhelming vote (50%) signals a governance crisis severe enough to threaten a board spill.

Tables 16 and 17 present the historical frequency data that calibrates the Board utility function's spill-risk component and contextualises the severity of the Qantas 2023 vote.

6.4.7 Board Overconfidence Bias on the Vote

When the Board is the focal actor, two cognitive biases distort its perception of the vote distribution:

Year	ASX 200	ASX 300	Avg Against (%)	Context
2020	∼22	∼40	∼34	Post-COVID; JobKeeper sensitivity
2021	∼28	∼35	∼35	"War for talent" concerns
2022	22	∼35	34.2	Return to normalcy; retention grant disputes
2023	23	**41 (record)**	**45.7**	Cost-of-living backlash; record severity

Table 16: First strikes on ASX remuneration reports by year, 2020–2023. The 2023 season recorded the highest number and average severity since the two-strikes rule was introduced in 2011.

Company	Year	Prior Yr (%)	Latest (%)	Corrective Action	Root Cause
Cromwell Property	2020	>25	>25	No (defensive)	Board disputes + strategy dissent
Crown Resorts	2021	>25	>25	Yes (board renewal)	Governance failures + regulatory
Westpac	2021	>25	>25	Yes (risk overhaul)	Regulatory lawsuits + fraud
Goodman Group	2022	>25	>25	Yes (10% LTI cut)	LTI quantum misalignment
Lovisa	2022	>25	33.0	No (resistant)	CEO sign-on bonuses
Dicker Data	2022	>25	33.0	No (resistant)	Pay-performance misalignment
Link Administration	2022	>25	31.0	Yes (improved disclosure)	Poor STI/LTI transparency
Lake Resources	2023	34.8	50.0	Yes (malus + clawback)	STI transparency + share price
IDP Education	2023	>25	>25	No (insufficient)	High incentives despite weak price
Reliance Worldwide	2023	>25	38.9	No (resistant)	Pay vs performance alignment
Aust. Clinical Labs	2023	>25	>25	Yes (cyber uplift)	Safety governance + exec pay

Table 17: Second-strike cases with board responses, 2020–2023. No incumbent director has ever lost their board seat at a spill meeting since the two-strikes rule was introduced in 2011; spill resolutions averaged <7% support in 2024 and 4.5% in 2025. The mechanism functions primarily as a communication tool and pressure valve, not a genuine removal threat.

Overestimation of governance effectiveness. The Board overestimates how much its actions reduce protest by a factor $\beta \sim U(0.25, 1.0)$ (production midpoint $\beta = 0.625$):

$$f_{\text{bias}}(D_1 = \text{review}) \sim U(0.63, 1.0), \qquad f_{\text{bias}}(D_1 = \text{CEO exit}) \sim U(-0.62, 0.5) \tag{36}$$

compared to the unbiased $U(0, 1)$ and $U(-1, 0.5)$ respectively. The biased bounds (in particular, the ceiling of 0.5 on the CEO-exit governance effect, which allows the Board to believe forced transition actually helps its position slightly) reflect the well-documented "illusion of control" bias in corporate governance, whereby boards systematically overestimate the efficacy of their own interventions (Malmendier & Tate, 2008). The specific bound values were calibrated to match the magnitude of documented board forecast errors in comparable Australian corporate crises.

Overprecision on vote uncertainty. The Board perceives $\kappa \sim U(2, 5)$ times more precision than warranted:

$$\hat{\sigma}_V = \frac{\sigma_V}{\sqrt{\kappa}} \tag{37}$$

At the production default $\kappa = 3.5$, the Board's subjective confidence intervals are roughly half the width of the true intervals (a σ_{scale} of approximately 0.53), causing it to underestimate the probability of extreme vote outcomes. This parameterisation is grounded in a specific finding from the overconfidence literature: Moore and Healy (2008) identify *overprecision* (the excessive certainty that one's estimates are accurate) as the most robust and persistent form of overconfidence, more reliable across contexts than either overestimation or overplacement. While the original Malmendier and Tate (2005) evidence was developed for CEO overconfidence in investment decisions, subsequent work (Malmendier & Tate, 2008) extended the framework to board-level strategic decision-making, where the same cognitive mechanisms apply: boards that have overseen sustained financial performance develop calibrated confidence in their own judgement that transfers poorly to novel crisis settings. A σ_{scale} of 0.53 implies the Board perceives roughly half the true uncertainty in vote outcomes, a compression consistent with the empirical magnitudes documented in the overprecision literature and calibrated to the Qantas-specific context where a decade of stable 8–10% remuneration votes provided no signal of the 82.9% escalation that actually materialised.

Both biases propagate consistently: the Board acts on its biased beliefs, and opponents model the Board as holding those beliefs, producing self-consistent decision-making under cognitive bias. The parameterisation is research-grounded in its framework, with specific numerical values involving judgement in mapping from the literature to the Qantas game tree.

6.4.8 Monte Carlo Integration at the Vote Node

At the V node in the game tree, the evaluator performs Monte Carlo integration:

1. Draw the governance effect f once (epistemic) and the crisis floor V_{floor} once (epistemic).

2. Draw $M_V = 50$ vote samples (aleatoric): for each sample m, construct B_{agm}, draw $\text{logit}(V_m) \sim \mathcal{N}(\alpha_V + B_{\text{agm}}, \sigma_V)$, apply expit, apply the floor if applicable, and derive the strike and overwhelming indicators.

3. Average downstream utility across all samples:

$$U_i^{(V)}(h) = \frac{1}{M_V} \sum_{m=1}^{M_V} U_i^{(\text{next})}(h \cup \{V_m, \text{strike}_m, \text{overwhelming}_m\}) \tag{38}$$

6.5 Modelling the CEO

The CEO has one decision, which may be **conditionally repeated**: whether to resign. At D_0^{ceo} (pre-game), the CEO chooses to resign or stay. If the CEO stays, a second decision arises at D_4 (post-AGM) and potentially D_4' (post-review): stay, resign, or negotiate an exit.

> *Note:* Theoretically, a further decision exists: whether to show contrition, take responsibility, and position himself as the solution to the problem. However, based on the CEO's entire history of public behaviour, this outcome was assessed as so unlikely that we decided to leave it out of the modelling.

Unlike the Board and ASA, the CEO's utility function is **literature-calibrated** rather than estimated via an LLM-and-Stan pipeline. This reflects two constraints: (1) the CEO made a single binary choice (resign on 5 September 2023), which is insufficient to identify a multi-parameter utility function econometrically; and (2) Joyce's compensation structure, clawback provisions, contract terms, and departure circumstances are publicly documented in sufficient detail to support credible calibration from established behavioural economics without statistical estimation.

6.5.1 Reference-Dependent CRRA Utility

The CEO's total utility is:

$$u_C(Z) = U_{\text{money}}(W) - \lambda_D \cdot D_{\text{raw}} \tag{39}$$

The monetary component uses constant relative risk aversion (CRRA) with Kahneman–Tversky loss aversion around a reference point W_{ref}:

$$\text{CRRA}(W) = \frac{W^{1-\gamma}}{1-\gamma}, \quad \gamma \neq 1 \tag{40}$$

$$U_{\text{money}}(W) = \begin{cases} \text{CRRA}(W) & W \geq W_{\text{ref}} \\ \lambda \cdot \text{CRRA}(W) - (\lambda - 1) \cdot \text{CRRA}(W_{\text{ref}}) & W < W_{\text{ref}} \end{cases} \tag{41}$$

This is continuous at W_{ref} and amplifies the utility drop below the reference point by exactly λ. In practice, all game outcomes have $W < W_{\text{ref}}$, so the CEO is always in the loss domain.

The non-monetary component D_{raw} captures reputational damage, public humiliation, career destruction, and legal exposure. It is additive across game outcomes and scaled by a separate loss aversion coefficient λ_D:

$$\begin{aligned} D_{\text{raw}} = D_{\text{base}} &+ D_{\text{departure}}(\tilde{d}_4) \cdot \mathbf{1}[\text{CEO\_removed}] \\ &+ D_{\text{agm}} \cdot \mathbf{1}[V > 0.25] + D_{\text{disgrace}} \cdot \mathbf{1}[\text{overwhelming}] \\ &+ D_{\text{adverse}} \cdot \mathbf{1}[\text{review\_negative}] \end{aligned} \tag{42}$$

where $\tilde{d}_4$ is the effective CEO departure action (d_4' overrides d_4 if the CEO acted post-review).

6.5.2 Preference Parameters

Parameter	Value	Source	Description
γ	1.5	Tversky and Kahneman (1992)	Risk aversion coefficient
λ	2.25	Tversky and Kahneman (1992)	Loss aversion (monetary)
W_{ref}	16.0	Pre-crisis expected comp. (A$M)	Reference point for loss aversion
λ_D	2.25	Default = λ	Loss aversion (non-monetary)

Table 18: CEO preference parameters.

The reference point $W_{\text{ref}} = 16.0$ reflects Joyce's pre-crisis expected total compensation (salary + STI + LTI) of approximately A$16M. His initial FY23 remuneration was reported at A$21.4M before clawbacks, and he executed a A$16.85M share sale in June 2023. With loss aversion $\lambda = 2.25$ from cumulative prospect theory (Tversky & Kahneman, 1992), losses below W_{ref} are amplified by a factor of 2.25. Malmendier and Tate (2005) document that overconfident CEOs with high media profiles exhibit stronger loss aversion, supporting a value at or above the population estimate for Joyce specifically.

6.5.3 Wealth Outcomes

Wealth depends on the CEO's departure mode. All values reflect post-ACCC calibration: the Board had flagged clawback of up to A$14.4M, LTIs were frozen, STIs were under scrutiny, and legal costs were mounting.

Departure mode	Parameter	W (A$M)	Rationale
Pre-game resign (D_0)	W_{resign}	8.0	Good leaver status; partial bonus; moderate clawback
Stay & kept	$W_{\text{stay\_kept}}$	7.0	Frozen LTIs, reduced STI, legal costs
Stay & negotiate (D_4)	$W_{\text{negotiate}}$	3.75	Midpoint of sacked and kept
Stay & resign late (D_4)	$W_{\text{resign\_late}}$	1.95	$1.3 \times W_{\text{sacked}}$; post-AGM timing
Stay & sacked	$W_{\text{stay\_sacked}}$	0.5	Near-total forfeiture after forced removal

Table 19: CEO wealth outcomes by departure mode.

A key insight: even the best stay outcome ($W = 7.0$) is below $W_{\text{ref}} = 16.0$. Combined with loss aversion, the monetary component alone slightly favours resignation. The non-monetary penalties then amplify this asymmetry.

6.5.4 Non-Monetary Penalties

The resign path produces $D_{\text{raw}} = D_{\text{resign}} = 40$. The worst-case stay path produces $D_{\text{raw}} = 25 + 100 + 30 + 30 + 10 = 195$ (sacked after overwhelming vote and adverse review). With $\lambda_D = 2.25$: effective penalty $= 2.25 \times 195 = 438.75$, compared to the resign path $2.25 \times 40 = 90$.

Parameter	Value	Condition	Calibration anchor
D_{stay}	25	CEO stays (baseline)	Ongoing ACCC exposure, hostile media
D_{resign}	40	CEO resigns at D_0	Karpoff, Lee, and Martin (2008): 70–90% turnover post-enforcement
D_{sacked}	100	Board fires CEO	Maximum reputational destruction (cf. AMP, Crown)
$D_{\text{resign\_late}}$	60	CEO resigns at D_4	Controls narrative but damage done (cf. Rio Tinto)
$D_{\text{negotiate}}$	45	Negotiates exit at D_4	Face-saving terms (cf. AMP 2021)
D_{agm}	30	Vote $>$ 25% (strike)	Public rejection by shareholders
D_{disgrace}	30	Overwhelming vote	Additional public disgrace
D_{adverse}	10	Negative review	Adds to stigma

Table 20: CEO non-monetary penalty parameters.

6.5.5 Pre-Game Resignation: Bayesian Prior and Level-2 ARA

The D_0^{ceo} decision is anchored by an empirical Bayesian prior derived from ASX moral-reputational crisis events. The full dataset of 26 ASX 100 CEO crisis events is presented in Table 21. The overall departure rate is 19/26 (73.1%), but this unconditional rate includes CEOs who adopted visible contrition strategies. When the sample is partitioned by contrition strategy (Table 22), the conditional departure rate given no contrition is 12/12 (100%).

Starting from a Jeffreys prior $\text{Beta}(0.5, 0.5)$ updated with the 12 no-contrition observations, all of whom departed:

$$p_{\text{departure}} \sim \text{Beta}(12.5,\ 0.5), \qquad \mathbb{E} = 0.962, \quad 90\%\ \text{CI} = [0.85,\ 0.998] \tag{43}$$

Joyce maps to the "no contrition" archetype: combative public posture throughout COVID complaints and ghost flights, record bonus while workers were stood down, lobbying against competition, and zero accountability signalling.

This prior is combined with a **Level-2 ARA prediction**. The CEO's resignation probability depends critically on how the model represents the Board's behaviour. A naive Level-1 approach treats the Board's future actions as fixed probabilities that do not respond to what the CEO decides. Under this assumption, the CEO evaluating "what happens if I stay?" averages expected utility across all three Board options including the relatively benign "do nothing" outcome, which carries enough weight to make staying appear tolerable. This produces a resignation probability of only around 35%, the opposite of what actually occurred.

The model instead requires the CEO to reason one step further: if I stay, what will the Board actually do *in response to the fact that I have stayed*? The answer, drawn from the Board's own decision model, is that commissioning a governance review dominates with 99% probability regardless of whether the CEO is present. The "do nothing" option that made staying attractive under the naive approach is effectively unavailable, because a strategically rational Board facing ACCC proceedings, a Senate inquiry, and a High

#	Company	CEO	Crisis Event	Status	Departure
1	CBA	Ian Narev	AUSTRAC AML scandal (Aug 2017)	Departed	Apr 2018
2	AMP	Craig Meller	Royal Commission: fees for no service (Apr 2018)	Departed	Apr 2018
3	NAB	Andrew Thorburn	Royal Commission: culture/remuneration (Feb 2019)	Departed	Feb 2019
4	Westpac	Brian Hartzer	AUSTRAC AML/child exploitation (Nov 2019)	Departed	Dec 2019
5	Rio Tinto	JS Jacques	Juukan Gorge heritage destruction (May 2020)	Departed	Jan 2021
6	Crown Resorts	Ken Barton	Bergin Inquiry: money laundering (Feb 2021)	Departed	Feb 2021
7	Star Entertain.	Matt Bekier	Bell Inquiry: CUP card/junket (Mar 2022)	Departed	Mar 2022
8	AMP	F. De Ferrari	Misconduct/cultural crisis (Apr 2021)	Departed	Dec 2021
9	James Hardie	Jack Truong	Workplace behaviour/misconduct (Jan 2022)	Departed	Jan 2022
10	QBE Insurance	Pat Regan	Workplace communication breach (Sep 2020)	Departed	Sep 2020
11	Tabcorp	Elmer Kupper	Regulatory investigation: Cambodia (Mar 2016)	Departed	Mar 2016
12	IOOF	Chris Kelaher	APRA legal action/Royal Commission (Dec 2018)	Departed	Apr 2019
13	Bapcor	Darryl Abotomey	Board governance clash (Nov 2021)	Departed	Dec 2021
14	ASX Ltd	Dominic Stevens	CHESS replacement failure (Nov 2022)	Departed	Aug 2022
15	Sigma Health	Mark Hooper	Major contract loss (FY 2017)	Departed	FY 2017
16	Bellamy's	Laura McBain	Chinese inventory collapse (Jan 2017)	Departed	Jan 2017
17	Lendlease	Steve McCann	Systemic underperformance (May 2021)	Departed	May 2021
18	Telstra	Andy Penn	Strategy stagnation/board pressure (May 2022)	Departed	Sep 2022
19	Treasury Wine	Michael Clarke	China tariff crisis/inventory glut (Jan 2020)	Departed	Jun 2020
20	BHP	Andrew Mackenzie	Samarco dam collapse in Brazil (Nov 2015)	Stayed	Left Jan 2020
21	ANZ	Shayne Elliott	Royal Commission fallout (Feb 2019)	Stayed	n/a
22	Woolworths	Brad Banducci	Systemic wage theft (Oct 2019)	Stayed	n/a
23	Domino's	Don Meij	Franchisee wage underpayment (Aug 2017)	Stayed	n/a
24	Medibank	David Koczkar	Massive customer data breach (Oct 2022)	Stayed	n/a
25	Link Group	Vivek Bhatia	UK FCA warning/Woodford redress (Sep 2022)	Stayed	n/a
26	Origin Energy	Frank Calabria	Takeover turmoil/transition (2022/23)	Stayed	n/a

Table 21: ASX 100 CEO departures after moral-reputational crises ($n = 26$). Overall departure rate: $19/26 = 73.1\%$. Cases 1–12 are the no-contrition departures that form the Beta(12.5, 0.5) prior; cases 20–23 are contrition survivors.

Court loss would never choose inaction. Staying therefore means near-certain review, elevated vote severity, possible adverse findings, and substantial risk of forced removal post-review.

The difference between these two approaches is not technical but conceptual. Governance crises are not situations where each party acts independently of the others. The Board responds to what the CEO does, and the CEO anticipates that response. Modelling this interdependence, which is precisely the contribution of Level-2 ARA (Rios Insua, Banks, & Rios, 2009), is what drives the resignation probability from 35% to 96.2% and produces a result consistent with what actually happened.

The engine combines the Beta prior with ARA-computed evidence via pseudo-count addition:

$$\Pr(\text{resign}) = \frac{12.5 + \sum_{i=1}^{N} \Pr(\text{resign} \mid \text{draw}_i)}{13.0 + N} \tag{44}$$

With $N = 100$ ARA draws, the prior has weight $13/(13 + 100) = 11.5\%$, a meaningful anchor from empirical data while allowing the game-theoretic analysis to dominate.

6.5.6 Post-AGM Decisions and Departure-Mode Resolution

At D_4 (post-AGM), the CEO chooses: stay, resign, or negotiate exit. The departure-mode penalty $D_{\text{departure}}(\tilde{d}_4)$ depends on the effective departure action:

$$D_{\text{departure}}(\tilde{d}_4) = \begin{cases} D_{\text{negotiate}} = 45 & \tilde{d}_4 = \text{negotiate\_exit} \\ D_{\text{resign\_late}} = 60 & \tilde{d}_4 = \text{resign} \\ D_{\text{sacked}} = 100 & \text{otherwise (Board fires CEO)} \end{cases} \tag{45}$$

	No Contrition	Contrition
Departed	12	0
Stayed	0	4

Table 22: Contrition strategy partition ($n = 16$, moral-reputational subset). Contrition is defined as voluntary incentive forfeiture + unreserved public apology + explicit personal ownership of remediation. The partition perfectly separates survivors from departures. No-contrition departures: Narev, Meller, Thorburn, Hartzer, Jacques, Barton, Bekier, De Ferrari, Truong, Regan, Kupper, Kelaher. Contrition survivors: Elliott (ANZ), Banducci (Woolworths), Meij (Domino's), Mackenzie (BHP).

Historical data shows the CEO is almost always replaced after a major ESG crisis. After any first strike, the Board's fixed policy sacks the CEO, so the CEO "sees" near-certain sacking risk. Typical D_4 predictive distribution after a strike: negotiate $\sim$72%, resign $\sim$26%, stay $\sim$2%.

6.5.7 Opponent Priors on CEO Parameters

When other actors model the CEO's behaviour, they sample CEO utility parameters from Normal prior distributions centred on the specification defaults (Table 18–20) with meaningful uncertainty. For example, the Board's prior on $D_{\text{sacked}} \sim \mathcal{N}(100, 30)$ reflects genuine uncertainty about how severely the CEO would experience a forced removal, while $W_{\text{resign}} \sim \mathcal{N}(8.0, 2.0)$ reflects uncertainty about the CEO's actual financial position. These opponent priors provide the uncertainty quantification that a Bayesian estimation pipeline would otherwise deliver.

6.6 Modelling the Governance Review

The final chance node in the game tree is R (Review), which resolves when the Board commissions an external governance review. Unlike the strategic actors modelled in Sections 5.2–5.5, the review is a **chance node**: its outcome is stochastic but not strategic. No utility function is required; only a calibrated probability model over the review's three output channels:

1. **Qualitative outcome rating** (negative, balanced, or positive), which triggers downstream feasibility rules (post-review round activation) and utility penalties.
2. **Cumulative abnormal return (CAR)**: a continuous market reaction to the findings release, entering the Board utility function via the review CAR weight.
3. **Direct costs**: reviewer fees, management distraction, and internal resource consumption.

Each component is modelled as a separate stochastic draw, calibrated from distinct empirical sources.

6.6.1 Empirical Calibration Data

The primary empirical basis is a longitudinal event study of external governance reviews for ASX-listed entities, using market-adjusted abnormal returns $AR_{i,t} = R_{i,t} - R_{m,t}$ with

CARs computed across three event windows (announcement, reviewer revealed, findings released). Six case studies provide calibration data:

Table 23: ASX Governance Review Case Studies (2014–2023)

Company	Period	Total CAR	Findings Window AR
CBA	2017–18	+0.93%	+1.75%
Westpac	2019–20	−7.30%	−3.00%
Rio Tinto	2020	−4.05%	−2.65%
Star Entertainment	2021–22	−19.15%	−13.95%
BOQ	2022–23	N/A	−5.70%
Qantas	2023–24	+0.20%	+0.85%

The findings window exhibits extreme heterogeneity: from +1.75% (CBA relief rally) to −13.95% (Star existential threat), motivating heavy-tailed distributional choices.

A broader panel of 21 ASX 100 governance reviews (2013–2023) is presented in Table 24. Three structural patterns emerge: regulatory reviews dominate (19 of 21 were regulator-initiated or prompted); negative outcomes predominate in regulatory reviews (all APRA governance self-assessments produced negative findings); and positive outcomes are rare, limited to proactive non-regulatory contexts.

Company	Period	Review Mechanism	Type	Outcome
CBA	2017–18	APRA Prudential Inquiry	Regulatory	Negative
Westpac	2020–24	Promontory Independent Review (CORE)	Regulatory	Negative
ANZ	2018–19	APRA Governance Self-Assessment	Regulatory	Negative
NAB	2018–19	APRA Governance Self-Assessment	Regulatory	Negative
Macquarie	2018–19	APRA Governance Self-Assessment	Regulatory	Negative
AMP	2019–21	APRA/ASIC Compliance & Governance	Regulatory	Negative
Rio Tinto	2020	Board Review of Cultural Heritage	Non-Regulatory	Negative
Crown Resorts	2020–21	Bergin Inquiry (NSW Casino Control Act)	Regulatory	Negative
Star Entertain.	2021–22	Bell Inquiry / Gotterson Inquiry	Regulatory	Negative
ASX Ltd	2018, 2023	Technology Governance & CHESS Inquiries	Regulatory	Negative
BOQ	2020	Independent Board Performance Review	Non-Regulatory	Positive
BOQ	2023	Remediation EU (APRA & AUSTRAC)	Regulatory	Negative
Suncorp	2018–19	APRA Governance Self-Assessment	Regulatory	Negative
IAG	2018–19	APRA Governance Self-Assessment	Regulatory	Negative
QBE	2018–19	APRA Governance Self-Assessment	Regulatory	Negative
Medibank	2018–19	APRA Governance Self-Assessment	Regulatory	Negative
Mineral Res.	2022	Governance Framework Gap Analysis	Non-Regulatory	Neutral
Telstra	2018–23	Compliance & Sales Practice Review	Regulatory	Negative
Woolworths	2020	Underpayment Governance Review	Regulatory	Negative
Tabcorp	2021–22	Regulatory Compliance Review	Regulatory	Negative
Bendigo Bank	2018–19	APRA Governance Self-Assessment	Regulatory	Negative

Table 24: ASX 100 external governance reviews, 2013–2023 ($n = 21$). Regulatory review outcome distribution: 19/20 = 95% negative. Positive outcomes are effectively absent in regulatory reviews and rare even in non-regulatory contexts.

6.6.2 Bayesian Posterior for the Qantas Context

The context-specific posterior updates the base rates using evidence specific to the September 2023 Qantas crisis: pre-existing reputational damage (ACCC ghost flights action, Senate inquiry), a 30% share price decline since March 2023, and the Board's incentive to signal accountability without creating litigation exposure.

Table 25: Bayesian Updating of Review Outcome Probabilities

Category	Prior	Posterior	Reasoning
Balanced	7–10%	75–85%	"Mistakes were made" is the dominant rational strategy for a board-initiated crisis review
Negative	3–5%	15–20%	Updated upward due to ACCC severity, but limited by board control of review scope
Positive	∼88%	<1%	A clean bill of health during active litigation would be non-credible

The dramatic posterior shift reflects the difference between the unconditional base rate (most board-initiated reviews are routine performance reviews that produce positive findings) and the conditional rate given crisis context.

6.6.3 Qualitative Outcome Rating: Dirichlet-Categorical Model

The outcome rating follows a two-level hierarchical model that separates epistemic and aleatoric uncertainty:

Level 1 (epistemic): Draw outcome probabilities once per belief draw:

$$(p_{\text{neg}}, p_{\text{bal}}, p_{\text{pos}}) \sim \text{Dirichlet}(38, 160, 1) \tag{46}$$

Level 2 (aleatoric): Each Monte Carlo sample draws:

$$\text{outcome} \sim \text{Categorical}(p_{\text{neg}}, p_{\text{bal}}, p_{\text{pos}}) \tag{47}$$

The Dirichlet is drawn once per belief draw because it represents epistemic uncertainty about the review process: the probabilities themselves are uncertain, not just which outcome occurs. Within a single belief draw, the probabilities are fixed and multiple MC samples draw from the same Categorical distribution.

Table 26: Dirichlet Pseudo-Count Calibration

Outcome	Pseudo-count	$\mathbb{E}[p]$	Interpretation
Negative	38	0.191	ACCC severity pushes well above 3–5% base rate
Balanced	160	0.804	"Mistakes were made" dominates as rational strategic path
Positive	1	0.005	Near-zero: clean bill of health non-credible during litigation

The total pseudo-count (199) controls concentration. At this level, the Dirichlet draws cluster tightly around the expected values, reflecting the strong posterior updating from Qantas-specific evidence. The pseudo-counts are derived by distributing the total concentration across categories in proportion to the posterior point estimates: $160 \approx 0.80 \times 199$ (balanced), $38 \approx 0.19 \times 199$ (negative), and 1 (minimum count ensuring positive outcomes remain in support).

When the Board is the focal actor, overconfidence inflates the positive pseudo-count:

$$\alpha_{\text{pos}}^{\text{biased}} = 1 \times (1 + 10 \times \beta_{\text{car}}) \tag{48}$$

With the default overconfidence parameter $\beta_{\text{car}} = 0.03$, this gives Dirichlet$(38, 160, 1.3)$, a slight tilt toward positive outcomes representing the Board's tendency to overestimate governance quality.

The outcome rating feeds into downstream game mechanics:

- Only a "negative" outcome triggers the post-review round ($D_4^{\text{post-review}}$ and $D_{\text{rev}}^{\text{post-review}}$).
- Board utility penalties fire unconditionally on CEO presence: $\delta_{\text{neg}} = 0.571$ for negative findings, $\delta_{\text{bal}} = 0.285$ for balanced.
- Positive findings carry no penalty.

6.6.4 Cumulative Abnormal Return: Hierarchical Student-t Model

The CAR from the findings release window follows a three-level hierarchy designed to accommodate the extreme heterogeneity observed in Table 23:

$$\mu_f \sim t(\nu = 4,\ -0.05,\ 0.03) \tag{49}$$

$$\sigma_f \sim \text{Half-Normal}(0.10) \tag{50}$$

$$\text{CAR} \sim t(\nu = 3,\ \mu_f,\ \sigma_f) \tag{51}$$

Location parameter (Equation 49): The $t(4)$ distribution is centred at -0.05 (-5%), reflecting a moderate negative expectation: governance reviews typically reveal bad news. The scale of 0.03 captures the range from CBA ($+1.75\%$) to Westpac (-3.00%), excluding the Star outlier, consistent with $\pm 2\sigma$ for a $t(4)$. The four degrees of freedom provide heavier tails than a normal while maintaining finite variance.

Scale parameter (Equation 50): The Half-Normal with scale 0.10 places most mass on $\sigma_f \in [0, 0.20]$, consistent with observed heterogeneity. A draw of $\sigma_f = 0.05$ produces CARs concentrated near μ_f; a draw of $\sigma_f = 0.15$ makes CARs of $\pm 20\%$ plausible.

Observation level (Equation 51): The $t(3)$ carries the heaviest tails in the hierarchy, with finite mean but infinite kurtosis. This ensures that "black swan" outcomes like Star's -13.95% remain in support. The $t(3)$ was chosen over thinner-tailed alternatives because the empirical distribution shows extreme negative kurtosis: one in six observations is a -14% outlier.

The CAR and outcome rating are modelled as **conditionally independent** given the belief draw. This is a deliberate simplification: the six-observation panel is insufficient to estimate conditional CAR distributions per outcome category. Moreover, the CAR reflects the *surprise* component relative to market expectations, since a balanced finding can produce a positive CAR if the market feared worse (as occurred in the actual Qantas case: balanced finding, $+0.85\%$ AR).

6.6.5 Direct Cost Model

Direct costs are modelled via:

$$C_{\text{direct}} \sim \text{Gamma}(\alpha = 4.55,\ \beta = 4741) \tag{52}$$

where C_{direct} is expressed as a CAR-equivalent decimal. The parameters are method-of-moments estimates from a first-principles cost decomposition calibrated for a reference market capitalisation of AUD 10 billion:

Table 27: Direct Cost Decomposition (CAR-equivalent basis points)

Component	Low	Central	High
Reviewer fees	2.0 bps	3.0 bps	5.0 bps
Management distraction	1.5 bps	4.0 bps	10.0 bps
Internal resources	1.2 bps	2.6 bps	4.5 bps
Total	**4.7 bps**	**9.6 bps**	**19.5 bps**

The resulting Gamma distribution has mean 9.6 bps, standard deviation 4.5 bps, and positive skewness (0.94), reflecting asymmetric risk: management distraction can escalate if the review becomes prolonged, but there is a natural floor on costs from reviewer fees alone. Direct costs ($\sim$10 bps) are small relative to the expected CAR from findings release ($\mathbb{E} = -500$ bps), ensuring they rarely serve as the binding constraint on the Board's review decision while preventing the model from treating reviews as costless.

6.6.6 Sampling Protocol and Game Tree Integration

At the R chance node, the tree evaluator executes a two-phase sampling protocol:

Phase 1 (Epistemic draw, once per belief draw):

- Draw outcome probabilities $(p_{\text{neg}}, p_{\text{bal}}, p_{\text{pos}}) \sim \text{Dirichlet}(38, 160, 1)$.
- These probabilities are held fixed across all MC samples within this belief draw.

Phase 2 (Aleatoric draws, per MC sample):

- Draw outcome $\sim \text{Categorical}(p_{\text{neg}}, p_{\text{bal}}, p_{\text{pos}})$.
- Draw CAR from the hierarchical Student-t (Equations 49–51).
- Draw direct cost $C_{\text{direct}} \sim \text{Gamma}(4.55, 4741)$.
- Apply state transition: if outcome = "negative" and CEO present, set post-review round active.

The review outcomes enter the Board utility function through three channels: the CAR impact (weight 15.0, multiplying the sampled CAR), the direct cost penalty (weight 15.0, multiplying C_{direct}), and discrete finding penalties ($\delta_{\text{neg}} = 0.571$ for negative, $\delta_{\text{bal}} = 0.285$ for balanced). The finding penalties are unconditional on CEO presence because review findings reflect on Board governance quality regardless of whether the CEO has departed.

6.6.7 Validation Against the Actual Outcome

The Qantas review was conducted by Tom Saar over approximately 10 months (October 2023 to August 2024), producing 32 recommendations. The actual outcome provides a single-point validation:

The balanced outcome matched the modal prediction with $\sim$80% posterior probability. The positive CAR (+0.85%) falls within the support of the $t(3)$ distribution though above the expected value, consistent with the market having priced in worse outcomes prior to findings release (the "relief rally" interpretation). The review also triggered a \$9.26 million clawback of the former CEO's payout and a 12-point reputation score uplift by mid-2024, consequences that flow through the game tree's post-review mechanics.

Table 28: Review Model Validation

Component	Model Prediction	Actual
Outcome rating	Balanced ($\mathbb{E}[p] = 0.804$)	Balanced
CAR	$\mathbb{E} = -5\%$ (wide uncertainty)	$+0.85\%$ (within $t(3)$ support)
Direct cost	$\mathbb{E} = 9.6$ bps	Not publicly disclosed

6.7 Replication Blueprint

The end-to-end workflow illustrated in Figure 1 can be summarised as a minimal implementation sequence for applying GenAI-assisted adversarial risk analysis to a new reputation-risk setting:

1. **Define the focal decision-maker and stakeholders.** Identify who has decision power, who influences outcomes indirectly, and who can be treated as exogenous context (Section 4.7).

2. **Gather institutional and public evidence.** Collect governance documents, regulatory filings, media commentary, and any available quantitative data (voting records, market reactions, prior incidents) for each strategic actor.

3. **Create stakeholder summaries and prompt packs.** Distil the evidence into structured persona descriptions and counterfactual decision scenarios suitable for LLM elicitation (Section 5.3).

4. **Run digital twin scenarios.** Query each stakeholder twin across the scenario set with repeated sampling and temperature diversity to produce distributional preference data (Section 5.4).

5. **Anchor outputs with historical data where available.** Use historical base rates, voting panels, and precedent cases to constrain digital twin outputs via informative priors and calibration targets.

6. **Estimate stakeholder models.** Fit the appropriate statistical model for each actor (ordinal probit for the Board, random-utility for the ASA, logit-normal for the vote, CRRA with opponent priors for the CEO), using the anchored digital twin data as input.

7. **Build the recursive game tree.** Specify the sequential decision structure, connect stakeholder models at each node, and implement the ARA recursion with sequential Monte Carlo (Section 6.1).

8. **Simulate, compare strategies, and report outputs.** Run the full simulation, compute expected utilities by action, generate counterfactual comparisons, sensitivity analysis, and value-of-information diagnostics.

This sequence is deliberately generic: while the specific model forms will vary by application, the workflow, from evidence gathering through digital twin elicitation to anchored Bayesian estimation and recursive evaluation, transfers directly to other reputation-risk settings.

6.8 Modelling Architecture Summary

Table 29 consolidates the modelling architecture across all five stakeholder components developed in this section. It complements the institutional architecture summary in Table 1 by focusing on the statistical model form, estimation method, decision rule, and key parameter sources for each actor.

Actor	Model form	Estimation method	Decision rule	Key parameter sources	Section
Board	Ordinal probit over utility weights	Stan MCMC (Bayesian)	Argmax-count over posterior draws	3,800 LLM scenario responses; 26 ASX 100 crisis cases	6.2
ASA	Random utility over context and interaction weights	Stan MCMC + constrained optimisation on elicited ranges	EU maximisation via random utility; Beta opponent prior when non-focal	LLM range/gap elicitation; 15 comparable governance incidents	6.3
Shareholders	Logit-normal vote with AR(1) belief dynamics	State-space SMC; 36-company ASX panel for mobilisation effects	Nature node	Qantas AGM vote history 2020–2023; cross-company panel of 36 recommendations	6.4
CEO	CRRA utility with opponent-prior beliefs	Method-of-moments on historical departures; literature-calibrated risk aversion	Prospect-theory expected-utility maximisation	12/12 no-contrition ASX CEO departures; Tversky and Kahneman (1992)	6.5
Governance Review	Dirichlet-categorical over outcomes; t-distributed CAR	Method-of-moments on historical review panel; Gamma direct-cost prior	Nature node	21 ASX governance reviews; 6-case CAR panel	6.6

Table 29: Modelling architecture for each stakeholder component: statistical model form, estimation method, decision rule, and key parameter sources. Model forms and decision rules complement the institutional roles and power channels documented in Table 1. Shareholders and the Governance Review are Nature nodes and have no decision rule.

7 Results

7.1 Posterior Predictive Check

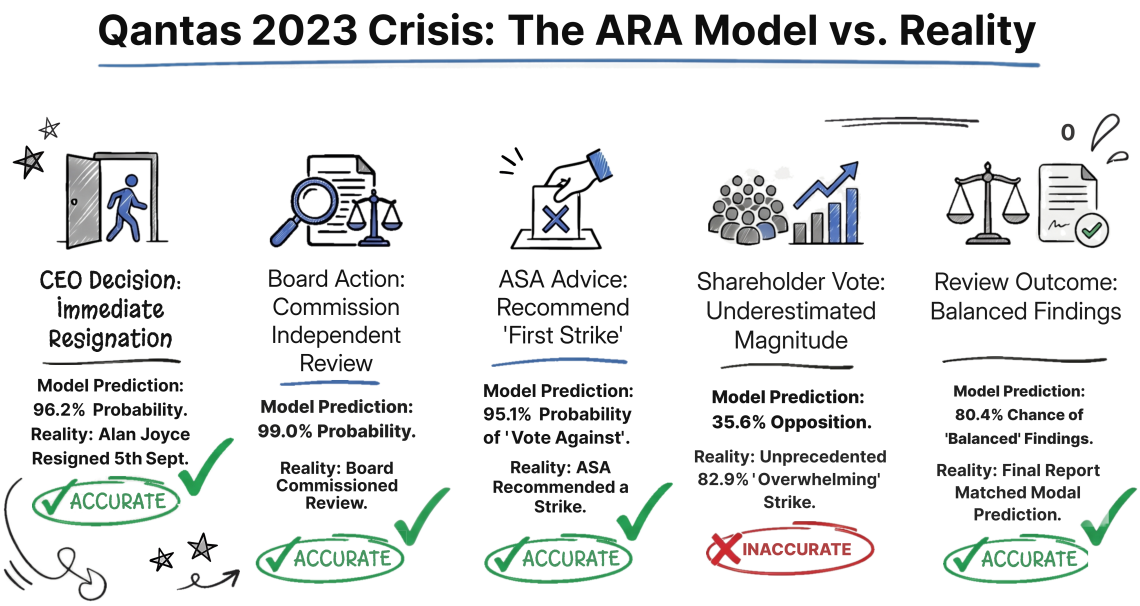

Figure 9: Timeline of the five key events in the 2023–2024 Qantas governance crisis, comparing the ARA model's *ex ante* predictions against the actual outcomes observed at each decision and chance node.

The most direct test of the ARA framework is whether it correctly identifies the sequence of observable events in the 2023–2024 Qantas governance crisis. Figure 9 summarises the comparison visually across the five key events, and Table 30 compares the model's predictions at each decision and chance node against what actually occurred.

Event	Predicted	Actual	Match	Notes
CEO departure (D_0^{ceo})	Pr(resign) = 96.2%	Resigned	✓	Beta(12.5, 0.5); 12/12 no-contrition departures
Board action (D_1)	Commission review (Pr = 99.0%)	Commission review	✓	Do nothing = 1.0%
ASA rec. (A_2)	Pr(strike) = 95.1%	Strike rec.	✓	Conditional on CEO resigned + D1_review
Vote fraction (V)	Mean 35.6%, SD 13.0%	82.9%	×	90% CI: [22.1%, 62.3%]; actual at 100th pctile
Review (R)	Pr(balanced) = 80.4%	Balanced	✓	Dirichlet(38, 160, 1); modal outcome

Table 30: Posterior predictive check: model predictions versus actual 2023–2024 outcomes. The model correctly identifies the modal outcome at four of five nodes. The vote fraction is the only miss: the model's prior predictive distribution, conditioned on the pre-crisis belief state, substantially underestimates the severity of the actual 82.9% vote.

The model correctly identifies the modal outcome at four of five nodes, including the two most consequential: CEO departure (96.2% predicted probability, outcome: resigned)

and Board action (99.0% predicted probability for commissioning a review, outcome: review commissioned). These are the nodes over which the Board retains genuine strategic discretion, and the model's recommendations at both nodes match what a well-governed board ultimately did. The ASA recommendation and governance review outcome are also correctly predicted. Four from five is a strong result for a model applied entirely prospectively, without any post-hoc tuning to the observed outcomes.

The single miss is the magnitude of the shareholder vote. The model's prior predictive mean was 35.6% (SD = 13.0%), while the actual vote was 82.9%, roughly 3.6 standard deviations above that mean. This is not a calibration error that better modelling could have prevented; it reflects a structural feature of the problem. The belief model is designed to track the gradual accumulation of reputational distrust, and performs well in that regime: the 2020–2022 votes of 8–10% are accurately captured, and the directional shift toward crisis is correctly identified. The 2023 result was driven by the simultaneous arrival of multiple regime-shifting events (ACCC Federal Court proceedings, the High Court outsourcing ruling, CEO resignation under pressure, and a Senate inquiry), producing a 73.6 percentage point escalation with no precedent in the data. No belief model anchored to prior votes could have anticipated that magnitude.

This is precisely the problem that motivates a Bayesian decision-tree approach. The question facing the Board was never "what will the exact vote be?" It was "given that the vote outcome is uncertain, including the possibility of outcomes well outside historical experience, what is the right thing to do now?" A point forecast, however sophisticated, cannot answer that question. A decision tree can, because it evaluates expected utility across the full distribution of possible outcomes, not just the most probable one. As the counterfactual analysis in Section 7.2 confirms, commissioning a governance review is the Board's optimal action whether the vote is 30% or 83%: the recommendation is robust to the very uncertainty the vote model cannot resolve. The 82.9% result does not undermine the framework; it illustrates why the framework is necessary.

7.2 Counterfactual Analysis

A posterior predictive check tells the analyst how well the model tracks what happened. Counterfactual analysis answers the more practically useful question: what should the Board have done, and does that recommendation hold up if history had unfolded differently? This is where the actuarial value of the framework is most directly demonstrated.

Table 31 presents Board expected utility under each feasible action at D_1, evaluated separately for the scenario that actually occurred (CEO resigned before the AGM) and the counterfactual in which Joyce remained in position. Two findings stand out.

The first is that commissioning a governance review is the optimal action in both branches. The Board did not need to know whether its CEO would resign to know what to do. That is not a trivial result: a model that recommended different actions depending on the CEO's unobservable intention would leave the Board paralysed, waiting for a signal that might never arrive clearly. Instead, the review is a dominant strategy, producing the highest expected utility whether the CEO is present or absent, and it is the action the Qantas Board actually chose. That alignment between the model's recommendation and the Board's eventual decision constitutes out-of-sample validation of the framework's practical relevance.

The second finding is arguably more important for practitioners. The EU gap

Scenario	D_1 Action	EU	Pr(strike)	Pr(overwh.)	$\mathbb{E}[V]$	Opt.
CEO resigned	Do nothing	−5.47	0.837	0.440	37.8%	
CEO resigned	**Comm. review**	**−4.94**	**0.831**	**0.435**	**35.7%**	✓
CEO stayed	Do nothing	+0.76	0.842	0.444	37.8%	
CEO stayed	**Comm. review**	**+1.69**	**0.837**	**0.440**	**35.7%**	✓
CEO stayed	Force CEO exit	+1.64	0.822	0.429	39.3%	

Table 31: Counterfactual analysis: Board expected utility under each D_1 action. Commissioning a governance review is the optimal action in both scenarios. In the CEO-resigned scenario, the review yields an EU advantage of 0.53 over inaction. In the CEO-stayed scenario, the review and forced CEO exit are nearly tied (EU difference of 0.05), both dominating inaction by ∼0.9.

between commissioning a review and forcing CEO exit is modest (0.05 in the CEO-stayed scenario), but the gap between either of those actions and doing nothing is substantial (approximately 0.9). In other words, the strategic choice facing the Board was not primarily *which* corrective action to take; it was *whether* to act at all. Inaction is the worst option by a wide margin regardless of what the CEO does. This quantifies something governance advisers often assert but rarely demonstrate: in a reputation crisis of this character, hesitation is not a neutral holding position. It carries a measurable expected utility cost, driven by the inaction penalties ($w_{\text{inact\_base}}$, $w_{\text{inact\_no\_rev}}$) that the review eliminates. The model puts a number on the cost of waiting, which is precisely the kind of input a board needs when the instinct under pressure is to defer.

7.3 Expected Utility Decomposition

To understand why the review dominates, Figure 10 and Table 32 decompose Board expected utility into its constituent penalties under each action. Two points deserve emphasis.

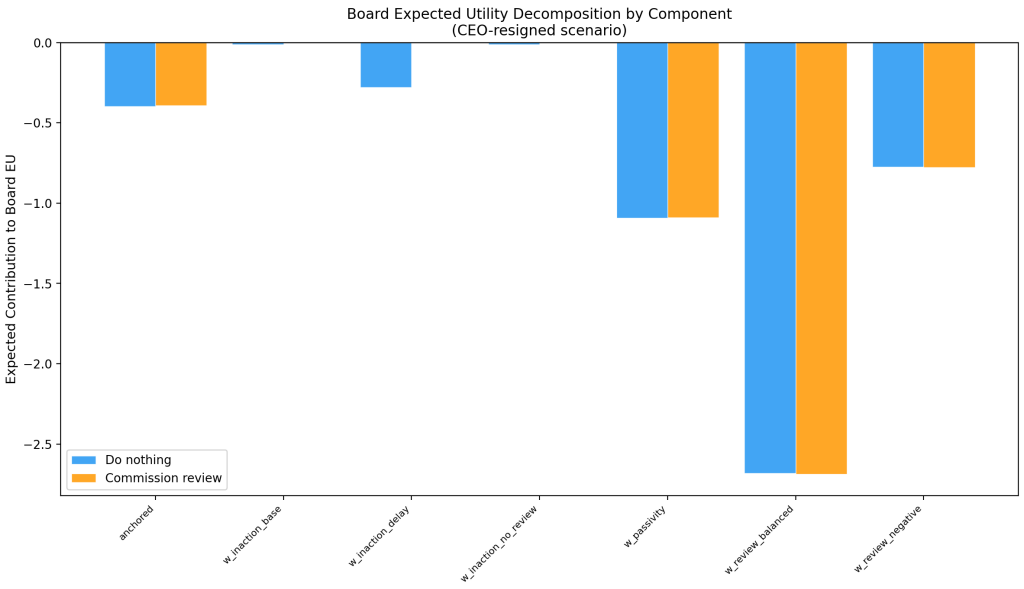

Figure 10: Board expected utility decomposition by component (CEO-resigned scenario). Blue bars show "Do nothing"; orange bars show "Commission review." The review eliminates the inaction-delay penalty ($w_{\text{inaction\_delay}}$) and the no-review penalty ($w_{\text{inaction\_no\_review}}$) at the cost of exposure to review-finding penalties ($w_{\text{review\_balanced}}$, $w_{\text{review\_negative}}$). The net effect favours the review.

There is no good option, only a less bad one. Both actions in Table 32 produce substantially negative expected utility (-5.47 for inaction, -4.94 for the review). A board looking at these numbers should not read them as a choice between pain and safety; they should read them as confirmation that by November 2023, significant governance pain was inevitable. The decomposition explains why: the $w_{\text{review\_balanced}}$ penalty (≈ -2.69) is the single largest term in *both* columns, driven by the near-certain probability of a "mistakes were made" finding. That penalty lands whether or not the Board commissions a review, because the governance reckoning is coming regardless of what the Board does. Choosing inaction does not avoid scrutiny; it simply adds the inaction penalties ($w_{\text{inact\_base}}$, $w_{\text{inact\_delay}}$, $w_{\text{inact\_no\_review}}$, totalling ≈ -0.30) on top of the same underlying exposure. The EU advantage of the review is real but modest precisely because the dominant cost is shared across both options.

The decomposition tells the actuary where to focus the client's attention. Vote penalties and CEO removal costs are negligible contributors to the EU gap. The Board's position is overwhelmingly determined by two terms: the balanced-finding penalty and the passivity penalty ($w_{\text{passivity}} \approx -1.09$). This has a direct practical implication. A board that responds to a crisis of this character by concentrating its energy on vote management or CEO transition optics is optimising the wrong variables. The model says clearly that the credibility, scope, and framing of the governance review and its findings is where the Board's leverage lies. Shifting the balanced-finding penalty by even a modest amount has a larger impact on expected utility than eliminating the vote and removal penalties entirely. That is the kind of prioritisation insight that justifies bringing an actuarial framework to a problem that boards typically navigate on instinct alone.

Component	Do nothing	Commission review
Anchored (CAR + costs)	−0.396	−0.392
$w_{\text{inaction\_base}}$	−0.012	n/a
$w_{\text{inaction\_delay}}$	−0.277	n/a
$w_{\text{inaction\_no\_review}}$	−0.012	n/a
$w_{\text{passivity}}$	−1.092	−1.090
$w_{\text{review\_balanced}}$	−2.682	−2.687
$w_{\text{review\_negative}}$	−0.775	−0.777
Total EU	**−5.468**	**−4.939**

Table 32: Expected utility decomposition by component (CEO-resigned scenario). The review eliminates three inaction penalties at the cost of marginally higher review-finding exposure. The net EU advantage is 0.53.

7.4 Sensitivity Analysis

Figure 11 presents a tornado chart showing the sensitivity of Board expected utility to ±50% variation in each utility weight parameter, evaluated in the CEO-stayed scenario at baseline EU = 1.648.

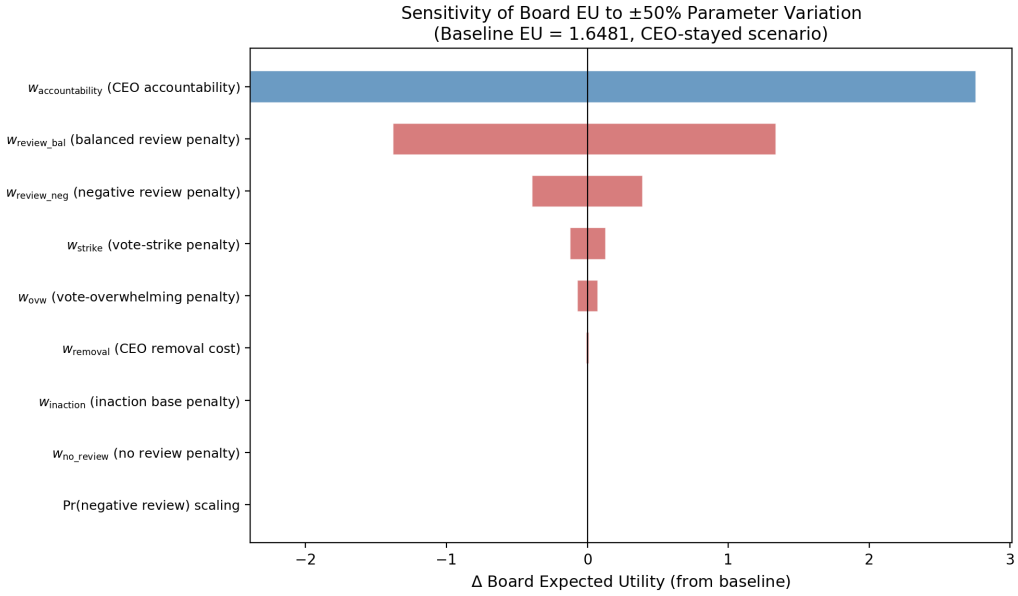

Figure 11: Tornado chart: sensitivity of Board EU to ±50% parameter variation (CEO-stayed scenario, baseline EU = 1.648). The CEO accountability weight ($w_{\text{accountability}}$) and balanced-review penalty ($w_{\text{review\_bal}}$) dominate; all other parameters have modest impact.

Table 33 provides the numerical values.

Two parameters dominate the sensitivity profile. The CEO accountability weight ($w_{\text{accountability}}$, swing = 5.15) captures the Board's benefit from removing a CEO who has demonstrably failed: halved, EU falls to −0.74; raised by 50%, it rises to 4.40. The balanced review penalty ($w_{\text{review\_bal}}$, swing = −2.72) is the second most influential term, because the model assigns ∼80% probability to a balanced finding and the penalty for that outcome scales directly with the weight. All six remaining parameters have swings

Parameter	EU (0.5×)	EU (baseline)	EU (1.5×)	Swing
$w_{\text{accountability}}$ (CEO accountability)	−0.743	1.648	4.403	5.146
$w_{\text{review\_bal}}$ (balanced review penalty)	2.983	1.648	0.267	−2.716
$w_{\text{review\_neg}}$ (negative review penalty)	2.039	1.648	1.256	−0.783
w_{strike} (vote-strike penalty)	1.774	1.648	1.523	−0.251
w_{ovw} (vote-overwhelming penalty)	1.720	1.648	1.576	−0.144
w_{removal} (CEO removal cost)	1.656	1.648	1.640	−0.015
w_{inaction} (inaction base penalty)	1.650	1.648	1.646	−0.003
$w_{\text{no\_review}}$ (no review penalty)	1.648	1.648	1.648	−0.001

Table 33: Sensitivity of Board EU to ±50% parameter variation (CEO-stayed scenario). The CEO accountability weight has the largest absolute swing (5.15); the balanced-review penalty has the second largest (−2.72). All other parameters have swings below 1.0.

below 1.0, and three of them below 0.02. Two conclusions follow from this profile, and both matter more than the point estimates themselves.

The recommendation is robust, and that robustness is itself the finding. The tornado chart is not just a sensitivity diagnostic; it is a stress test of the recommendation. The fact that commissioning a review remains optimal under every single-parameter ±50% perturbation means the Board does not need precise estimates of its own utility weights to act correctly. In a crisis setting where self-knowledge is compromised by pressure, groupthink, and cognitive bias, a recommendation that survives wide parameter uncertainty is qualitatively more valuable than one that is only optimal at a point estimate. The robustness result should be stated as a positive conclusion, not merely noted as the absence of a reversal: the review wins everywhere in the plausible parameter space, and that is the kind of result a board can act on with confidence.

The two dominant parameters tell you what the Board is actually afraid of, and whether that fear is well-founded. The sensitivity profile is highly concentrated: $w_{\text{accountability}}$ and $w_{\text{review\_bal}}$ together account for nearly all the variation in EU, while six other parameters are essentially inert. That concentration is informative in itself. It means the Board's expected utility is driven almost entirely by how much it values being seen to hold the CEO accountable, and how severely it weighs a balanced governance finding. A board that underweights accountability (low $w_{\text{accountability}}$) still finds the review optimal, but by a thin margin, which is exactly the cognitive risk profile of a board exhibiting the deference and passivity that the Qantas governance review later identified as a root cause of the crisis.

7.5 Vote Distribution by Board Action

Figure 12 shows the posterior predictive distribution of the shareholder vote fraction under each Board action, conditional on the ASA recommending a strike in a headline-incident scenario.

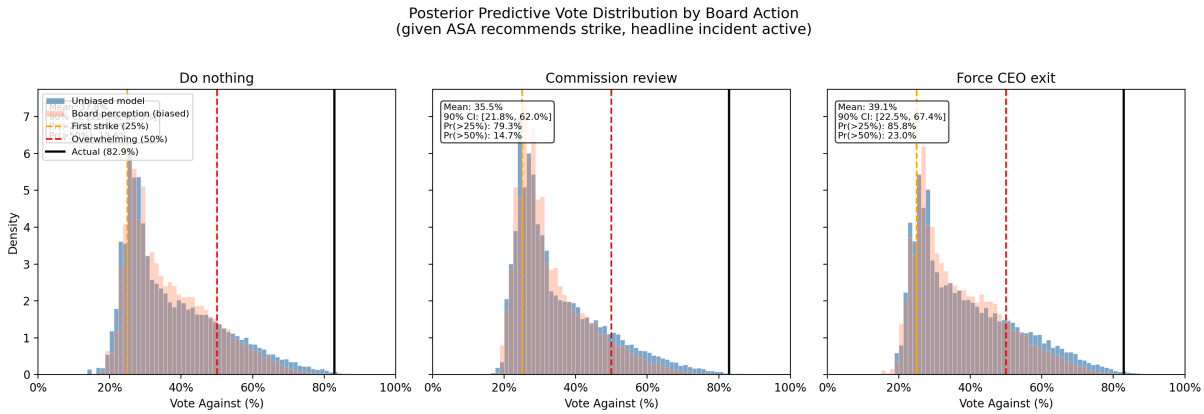

Figure 12: Posterior predictive vote distributions by Board action (given ASA recommends strike, headline incident active). Blue histograms show the unbiased model; orange shows the Board's biased perception. Dashed lines mark the 25% first-strike and 50% overwhelming thresholds; the solid black line marks the actual 82.9% vote. The model assigns 79–86% probability to a first strike and 15–23% probability to an overwhelming vote across all Board actions. The actual vote lies far in the right tail of all three distributions.

Three findings emerge from the vote distributions:

1. **The two-strikes rule had already lost its deterrent function.** The probability of exceeding the 25% first-strike threshold ranges from 79.3% under commission review to 85.8% under forced CEO exit; no available Board action can prevent a strike with meaningful probability. The strategic implication is that the Board's decision problem was never about avoiding a strike; it was about managing what came after one. A governance mechanism designed as a pre-commitment device to discipline boards had, by the time of the 2023 AGM, collapsed into a near-certain outcome regardless of Board behaviour, shifting the locus of strategic choice from prevention to response.

2. **The Board's overprecision bias is a quantitative account of why boards get blindsided.** The orange histograms are systematically more concentrated than the unbiased model (blue): with $\kappa = 3.5$, the Board's subjective confidence intervals are roughly half the true width. A board with systematically compressed intervals will consistently underestimate the probability of tail outcomes, which is precisely the tail outcome that actually materialised at 82.9%. This is not a modelling footnote; it is a measurable behavioural mechanism explaining why boards in governance crises so often appear blindsided by shareholder reactions they should have anticipated. The orange distribution in Figure 12 shows what the Board expected; the black line shows what occurred. The gap between them is the cost of overprecision.

3. **The actual 82.9% vote** lies far in the right tail of all three distributions, confirming the posterior predictive check finding that the model underestimates vote severity from pre-crisis priors. This underestimation is structural: the belief model is calibrated from 2020–2022 votes of 8–10%, and the 73.6 pp escalation exceeds any precedent in the training data.

The purpose of this comparison is not to tune the model ex post to match an unprecedented realised vote. Doing so would be misleading. The 82.9% result is informative precisely because it sits outside the range implied by prior crises and shows that governance escalation can exceed historical precedent. In that sense, the case study reinforces the value of a branching strategic model: the relevant question is not whether the exact vote can be forecast perfectly, but whether the decision framework still identifies robust early actions when extreme outcomes remain possible. As the counterfactual analysis in Section 7.2 confirms, commissioning a governance review is the optimal Board action across the full range of vote outcomes, including the unprecedented one that actually materialised.

7.6 Value of Information

In a governance crisis, the instinct to wait is almost universal. Boards delay action pending clarity on what proxy advisers will recommend, how institutional shareholders are leaning, and whether the situation will resolve itself without forcing a confrontation. That instinct feels prudent. The value of information analysis in Table 34 shows it is not, at least not in a crisis of this character.

Signal	**Baseline EU**	**EU with info**	**VoI**	**Detail**
ASA recommendation	−4.953	−4.934	0.019 (0.4%)	Optimal action is "commission review" regardless of ASA signal
Vote outcome	−4.953	−4.950	0.004 (0.1%)	Optimal action is "commission review" regardless of vote outcome

Table 34: Value of perfect information (VoI) for two key uncertainties. The VoI is negligible in both cases: the optimal Board action does not change under any realisation of the ASA recommendation or vote outcome.

The value of perfect information about the ASA's recommendation is 0.019 EU units, or 0.4% of baseline. The value of perfect information about the vote outcome is 0.004 EU units, or 0.1%. Both are negligible. Even if the Board could know with certainty what the ASA would recommend and exactly how shareholders would vote, it would not change what the Board should do. Commissioning a governance review is the optimal action under every realisation of both signals. There is no information the Board could acquire that would make inaction look better.

This is not a failure of the model to find interesting variation; it is one of its most useful outputs. The result is negligible precisely because the review's benefits, eliminating inaction penalties and signalling accountability, accrue independently of downstream vote outcomes, while its costs are modest relative to the penalties avoided. The Board's decision does not depend on how others will react because the review is justified across essentially all plausible states of the world.

The deeper point, and the one that transfers most directly to other reputation risk settings, is that this framework distinguishes between two fundamentally different types of crisis. In an *information-dominant* crisis, the right action depends critically on signals that are not yet available, and waiting to gather them has genuine strategic value. In an

action-dominant crisis, the recommendation is robust across the full range of plausible downstream states, and delay serves no strategic purpose while carrying a measurable expected utility cost. The Qantas case is firmly in the second category. Identifying which regime applies is itself a primary output of the ARA framework, and a more valuable one than the specific EU numbers, because it tells the decision-maker not just what to do but how urgently to do it.

For practitioners, the practical lesson is direct: when VoI is negligible, the board's priority should be speed of response rather than intelligence gathering. Pre-AGM engagement with shareholders and proxy advisers remains worthwhile for managing the *magnitude* of the vote, but it should run in parallel with decisive action, not as a precondition for it.

8 Limitations

Several limitations of this analysis should be acknowledged before drawing general conclusions.

Single case study design. The framework is demonstrated on one company, one crisis, and one AGM cycle. The Qantas case was chosen because the outcome is known and the institutional evidence is unusually rich, but this means the model's calibration choices (particularly the utility weight priors and the structural crisis floor) are tailored to a specific context. Whether the pipeline generalises to other sectors, regulatory environments, or governance structures remains an open empirical question. The replication blueprint in Section 6.7 is intended to make this testable, but external validation across multiple cases is needed before the approach can be treated as a general-purpose tool.

Reliance on a single LLM as elicitation engine. All digital twin elicitation uses gpt-4o-mini. This introduces two concerns. First, the model's training distribution shapes its simulated stakeholder responses in ways that are not fully observable; if the training data over-represents certain governance archetypes or under-represents Australian institutional contexts, the elicited preferences will be systematically biased in ways that are difficult to detect. Second, LLM alignment training may suppress extreme or combative responses, producing digital twins that are more conciliatory than their real-world counterparts. The repeated sampling and historical anchoring procedures described in Sections 5.3 and 5.4 partially mitigate these concerns but do not eliminate them. Replication using alternative models would strengthen confidence in the elicited preference structures.

Stability of the Board utility function. The ordinal probit estimation assumes the Board's utility weights are stable across the game horizon, from the pre-AGM period through to the post-review round. In practice, the governance review findings, the ACCC settlement, and the change in Board composition that followed Joyce's departure almost certainly shifted the weights, particularly the relative importance of financial performance versus stakeholder trust. The model captures a snapshot of Board preferences at a single crisis point rather than a dynamic preference structure that evolves as the crisis unfolds. Extensions that allow utility weights to update at key checkpoints (for example

at the ACCC announcement or the AGM itself) would better represent the Board's actual decision environment.

Vote magnitude underestimation. As discussed in Section 7.1, the model substantially underestimates the severity of the actual 82.9% vote. The belief model is designed to track gradual accumulation of reputational distrust and performs well in that regime. The 2023 Qantas vote reflects a discontinuous escalation driven by simultaneous regime-shifting events with no precedent in the training data. This is a genuine boundary condition of the approach rather than a calibration error, and no belief model anchored to prior votes could have anticipated this magnitude. The practical implication is that in crises of this character the model's contribution shifts from magnitude prediction to structural recommendation, which, as the counterfactual analysis confirms, remains valid across the full range of plausible vote outcomes including the unprecedented one that materialised.

9 Conclusion

This case study illustrates several key lessons for actuaries seeking to apply adversarial risk analysis to reputation risk:

1. **Adversarial risk analysis can be surprisingly complex.** Even a relatively contained scenario (a single company, five stakeholder groups, a handful of decisions) generates a large decision tree with many interacting uncertainties. The temptation to add detail and realism must be balanced against tractability and interpretability. **Don't overcomplicate things.**

2. **Start simple.** Begin with the most important decisions and the most influential stakeholders. Use simple conditional probability models where full utility elicitation is not feasible. Add complexity only where it materially changes the recommendation. A model that is transparent and approximately right is more useful than one that is precise and opaque.

3. **Boost the process with GenAI digital twins.** Large language models provide a transformative capability for adversarial risk analysis. Digital twins (generative AI agents calibrated to the public behaviour, stated values, and revealed preferences of real stakeholders) allow the analyst to:

    - Elicit preferences at scale and across scenarios that would be impractical to test with real individuals.
    - Stress-test models by simulating stakeholder responses to novel or extreme events.
    - Fill data gaps where historical precedent is sparse or where direct access to the decision-maker is unavailable.
    - Provide an auditable, reproducible record of the assumptions underlying each stakeholder model.

4. **The framework delivers robust recommendations under parameter uncertainty.** Because the recommendation emerges from expected-utility

comparison across the full posterior rather than from a point estimate, an ARA pipeline can produce an action that remains optimal across wide perturbations of the elicited utility weights, which is the property actuaries should be selling to clients whose self-knowledge is inevitably imperfect under crisis conditions.

5. **The framework diagnoses action-dominant versus information-dominant crises.** By quantifying the value of perfect information about downstream signals, ARA tells the decision-maker whether the priority should be speed of response or intelligence gathering, a regime distinction that is often asserted in governance practice but rarely quantified.

One of the model's most practically useful outputs is that the Board's optimal action, commissioning a governance review, is *action-dominant*: it produces the highest expected utility regardless of the ASA's recommendation, regardless of the shareholder vote outcome, and regardless of whether the CEO has resigned or remains in position. The value of waiting for confirming signals is negligible. In governance crises, decision-makers often delay action pending stakeholder reactions, hoping that the situation will resolve itself or that better information will change the calculus. The model shows that in this case, waiting had little strategic value because the review was justified across almost all plausible downstream states. Distinguishing between action-dominant and information-dominant crises is itself a valuable contribution of the framework: when early action dominates, the Board's priority should be speed of response rather than intelligence gathering.

Generative AI does not replace actuarial judgement; it extends the reach of actuarial methods into domains that were previously data-poor and model-resistant. Reputation risk is one such domain, and the combination of adversarial risk analysis with GenAI tools makes it a quantitative actuarial practice area for the first time.

9.1 Ethical Obligations and the Limits of Utility Maximisation

An actuary advising a board on reputation risk operates under the profession's Code of Professional Conduct, which requires that advice serve the public interest and not merely the client's private objectives. This obligation creates a tension with the ARA framework as presented: the model optimises the Board's expected utility, but the Board's fiduciary duties extend beyond its own governance exposure.

Consider a hypothetical illustration. Suppose the model identifies that the Board's expected utility is marginally higher if the governance review is commissioned but its terms of reference are narrowly scoped to exclude examination of the ghost flights conduct, on the grounds that broader scope increases the probability of a negative finding and triggers higher vote severity and greater CAR loss. The ARA framework, optimising solely over the Board's utility function, would recommend this narrow scoping because it reduces expected penalties without materially changing the review's signal value to shareholders.

This recommendation would be ethically and legally problematic on several grounds. The Board's statutory duty under section 181 of the Corporations Act requires directors to act in good faith in the best interests of the company, not merely in the interests of the current directors' tenure or reputation. A deliberately narrow review that shields the most publicly contested conduct from scrutiny would likely breach that duty. It would also expose individual directors to personal ASIC liability if the scoping decision were later characterised as designed to obstruct accountability. And it would almost certainly fail the

pub test: a review visibly designed to avoid the issue that triggered public outrage would be reported as a cover-up rather than a reckoning, compounding reputational damage rather than mitigating it.

The lesson for actuarial practice is that the utility function the model optimises is specified by the analyst, and a utility function that captures only the Board's short-term governance exposure omits the constraints that make some high-utility strategies illegitimate. The actuary's role is not simply to report the utility-maximising action but to flag when that action sits outside the feasible set defined by legal duties, professional ethics, and the reasonable expectations of the public. In adversarial risk analysis terms, the "pub test" is an additional constraint on the action space, not an afterthought.

9.2 Practical Takeaway

For practitioners, the immediate takeaway is straightforward: identify the stakeholders who actually possess decision power, distinguish them from background reputational forces, structure the strategic sequence explicitly, use GenAI to generate disciplined prior information where direct elicitation is impossible, and then constrain those priors with whatever empirical data and institutional evidence are available. That workflow will not eliminate uncertainty, but it allows actuaries to bring quantitative discipline to crises that would otherwise be analysed only through narrative instinct.

References

ABC News. (2022, August). *Qantas CEO Alan Joyce apology.* Retrieved from https://www.abc.net.au/news/2022-08-22/qantas-ceo-alan-joyce-apology/101356120

Actuaries Institute. (n.d.). *What is an actuary?* Retrieved from https://www.actuaries.asn.au/careers/what-is-an-actuary (Accessed 2024)

Aher, G. V., Arriaga, R. I., & Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. *Proceedings of the 40th International Conference on Machine Learning*, 337–371.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, *31*(3), 337–351.

ASX Corporate Governance Council. (2024). *Corporate governance principles and recommendations* (5th ed.; Tech. Rep.). ASX. Retrieved from https://www.asx.com.au/content/dam/asx/about/corporate-governance-council/reviews-and-submissions/2024/fifth-edition/australian-shareholders-association-submission.pdf

Australian Aviation. (2024, August). Too much deference to Joyce, says Qantas governance review. *Australian Aviation.* Retrieved from https://australianaviation.com.au/2024/08/too-much-deference-to-joyce-says-qantas-governance-review/

Australian Competition and Consumer Commission. (2025). *Domestic airline monitoring report: December 2025* (Tech. Rep.). Canberra: ACCC. Retrieved from https://www.accc.gov.au/system/files/domestic-airline-monitoring-report-december-2025.pdf

Australian Institute of Management. (2024). Qantas docks Joyce payout by $9.26m after board review finds failings. *AIM Knowledge Hub*. Retrieved from https://ami.org.au/knowledge-hub/

Australian Shareholders' Association. (n.d.-a). *Advocacy and monitoring.* Retrieved from https://www.australianshareholders.com.au/advocacy-monitoring/ (Accessed 2024)

Australian Shareholders' Association. (n.d.-b). *Focus issues and voting guidelines.* Retrieved from https://www.australianshareholders.com.au/focus-and-guidelines/ (Accessed 2024)

Australian Shareholders' Association. (2023). *Annual report 2023* (Tech. Rep.). Australian Shareholders' Association. Retrieved from https://www.australianshareholders.com.au/wp-content/uploads/2024/05/ASA-Annual-Report-2023_.pdf

Australian Shareholders' Association. (2025). *Focus issues 2025* (Tech. Rep.). Australian Shareholders' Association. Retrieved from https://www.australianshareholders.com.au/wp-content/uploads/2025/05/Focus-Issues-2025_MR.pdf

Banks, D. L., Rios, J., & Rios Insua, D. (2015). Adversarial risk analysis: An overview. *Wiley Interdisciplinary Reviews: Computational Statistics*, *7*(1), 35–45.

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, *120*(6), e2218523120.

Brand, J., Israeli, A., & Ngwe, D. (2023). Using GPT for market research. *Harvard Business School Working Paper*(23-062).

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... others (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Couce-Vieira, A., Insua, D. R., & Kosgodagan, A. (2020). Assessing and forecasting cybersecurity impacts. *Decision Analysis*, *17*(4), 356–374.

Ekin, T., Naveiro, R., Rios Insua, D., & Ruggeri, F. (2021). Augmented probability simulation methods for sequential games. *EURO Journal on Computational Optimization*, *9*, 100015.

Fischhoff, B. (1982). Debiasing. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 422–444). Cambridge University Press.

Grieves, M., & Vickers, J. (2017). Digital twin: Mitigating unpredictable, undesirable emergent behavior in complex systems. In *Transdisciplinary perspectives on complex systems* (pp. 85–113). Springer.

Horton, J. J. (2023). Large language models as simulated economic agents: What can we learn from Homo Silicus? *National Bureau of Economic Research Working Paper*(29631).

Intelligent Investor. (2023). Lessons from the ethical collapse of Qantas. *Intelligent Investor*. Retrieved from https://www.intelligentinvestor.com.au/investment-news/lessons-from-the-ethical-collapse-of-qantas/152898

International Actuarial Association. (2018). *Glossary of defined terms used in international standards of actuarial practice.* Retrieved from https://www.actuaries.org (Accessed 2024)

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–291.

Karpoff, J. M., Lee, D. S., & Martin, G. S. (2008). The cost to firms of cooking the

books. *Journal of Financial and Quantitative Analysis*, *43*(3), 581–612.
Larrick, R. P. (2004). Debiasing. *Blackwell Handbook of Judgment and Decision Making*, 316–338.
Malmendier, U., & Tate, G. (2005). Ceo overconfidence and corporate investment. *The Journal of Finance*, *60*(6), 2661–2700.
Malmendier, U., & Tate, G. (2008). Who makes acquisitions? CEO overconfidence and the market's reaction. *Journal of Financial Economics*, *89*(1), 20–43.
McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics*, 105–142.
Merrick, J. R. W., & Parnell, G. S. (2011). A comparative analysis of PRA and intelligent adversary methods for counterterrorism risk management. *Risk Analysis*, *31*(9), 1488–1510.
Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, *115*(2), 502–517.
Naveiro, R., Redondo, A., Rios Insua, D., & Ruggeri, F. (2019). Adversarial classification: An adversarial risk analysis approach. *International Journal of Approximate Reasoning*, *113*, 133–148.
Ownership Matters. (n.d.). *Regulation of proxy advisors: Our view.* Retrieved from `https://ownershipmatters.com.au/issues/regulation-of-proxy-advisers-our-view/` (Accessed 2024)
Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 1–22.
Qantas Airways. (2024). *Board governance review* (Tech. Rep.). Qantas Airways Limited. Retrieved from `https://www.qantas.com/au/en/about-us/our-company/governance.html`
Rios Insua, D., Banks, D. L., & Rios, J. (2009). *Adversarial risk analysis.* New York: Springer.
Rios Insua, D., Couce-Vieira, A., Rubio, J. A., Pieters, A., Wolter and"; Laszka, & Grossklags, J. (2021). An adversarial risk analysis framework for cybersecurity. *Risk Analysis*, *41*(1), 172–189.
Rios Insua, D., Rios, J., & Banks, D. L. (2009). Adversarial risk analysis. *Statistical Science*, *24*(2), 238–244.
Rios Insua, D., Rios, J., & Banks, D. L. (2015). Adversarial risk analysis for first-price sealed-bid auctions. *European Journal of Operational Research*, *244*(2), 567–578.
RMIT University. (2023, September). *Future Qantas.* Retrieved from `https://www.rmit.edu.au/news/media-releases-and-expert-comments/2023/sep/future-qantas`
Shanahan, M., McDonell, K., & Reynolds, L. (2023). Role play with large language models. *Nature*, *623*, 493–498.
The Sydney Morning Herald. (2024). Qantas governance and board culture. *The Sydney Morning Herald*.
Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.
Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297–323.
Walker, W. E., Lempert, R. J., & Kwakkel, J. H. (2013). Deep uncertainty. In S. I. Gass & M. C. Fu (Eds.), *Encyclopedia of operations research and management science* (3rd

ed., pp. 395–402). Boston, MA: Springer. doi: 10.1007/978-1-4419-1153-7_1140

Wang, S., & Rios Insua, D. (2019). Adversarial risk analysis for maritime piracy. *Decision Analysis*, *16*(3), 230–248.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., ... Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*, 24824–24837.

Westphal, J. D., & Zajac, E. J. (1997). Defections from the inner circle: Social exchange, reciprocity, and the diffusion of board independence in U.S. corporations. *Administrative Science Quarterly*, *42*(1), 161–183.

A Open-Source Code

All code, data pipelines, digital twin prompts, Stan models, and game-tree engine used in this paper are released as open source under the MIT licence. The repository is available at:

`https://github.com/colinpriest/qantas-adversarial-risk-analysis`

The repository includes the full elicitation pipeline, Bayesian estimation scripts, ARA decision engine, and result-generation notebooks required to reproduce every table and figure in this paper.