On the Identifiability of Retrospective Decision Error: A Bayesian and Counterfactual Inquiry into *Did You Fuck Up?*

Anonymous — Independent Research in Sequential Regret

October 2026

Abstract

We study retrospective decision error under incomplete records, stochastic outcomes, and costly opportunities for repair. A genuine fuck-up is defined by an avoidable deficit in ex ante expected utility, rather than by realized misfortune. We prove that observationally equivalent decision environments can disagree about whether an error occurred, and give conditions for consistent Bayesian diagnosis when informative evidence is available. A separate coupling argument shows that even identified marginal outcome distributions need not identify individual counterfactual benefit. A belief-state recovery model yields an intervention threshold, a finite-horizon stopping rule, and a synthesis theorem separating diagnosis from treatment. In a numerical example, learning that a mistake probably occurred makes intervention strictly less attractive. Information has value only through decisions it can improve; under conditionally uninformative rumination and strictly positive cost, the uniquely optimal policy is to go to bed. All named results are consequences of the stated models, not newly discovered laws of human behavior.

1 Preliminaries and the Decision Space

An individual chooses an action a_0 from a finite set A after observing an information set I_0. Examples include sending a risky text, remaining silent after a social misreading, or buying an asset whose downside has been carefully described in a document one has declined to open. Let $Y(a)$ denote the potential outcome under action a, and let u_0 be the utility criterion used to evaluate the original decision. Assume bounded utilities and a coherent original probability assessment.

An auditor may lack the original record. Let $\theta \in \Theta$, with Θ finite, index admissible reconstructions of the prior, likelihood, information actually available, preferences, and feasible actions. Define

$$q_\theta(a) = \mathbb{E}_\theta[u_0(Y(a)) \mid I_0], \qquad \Delta_\theta = \max_{a \in A} q_\theta(a) - q_\theta(a_0). \tag{1.1}$$

The notation suppresses the possible dependence of the archived record I_0 on θ. Crucially, θ does not authorize adding facts learned only after a_0. Decision quality is relative to this explicit benchmark; the model does not identify a uniquely correct prior or a universal utility function [1].

Definition 1.1 (Genuine fuck-up)**.** For a declared materiality tolerance $\varepsilon \geq 0$, the error event is

$$\mathsf{F} = \{\theta : \Delta_\theta > \varepsilon\}. \tag{1.2}$$

A bad outcome is the distinct event $B = \{u_0(Y(a_0)) < \ell\}$ for a fixed cutoff ℓ.

This is a decision-error definition, not a finding of moral blame. It presumes that the comparator actions were genuinely feasible, including their attention and acquisition costs. If gathering information was an available action, its cost belongs in q_θ; information that could not reasonably have been obtained is not silently placed in I_0.

2 The Fundamental Problem of Identifiability

Let D_n comprise the observed outcome and all evidence available to the auditor by time n. Write P_θ^n for its distribution. An error label is identifiable from these observations if equal observation laws imply equal labels: $P_\theta^n = P_{\theta'}^n$ for every n must imply $\mathbf{1}_{\mathsf{F}}(\theta) = \mathbf{1}_{\mathsf{F}}(\theta')$. This concerns an observation scheme, not the intensity with which one examines it.

Theorem 2.1 (The Bad Outcome Fallacy)**.** *For any $0 \le \varepsilon < 0.3$, there exist two decision models with identical distributions of every observed datum, but opposite error labels. For any randomized diagnostic rule $\delta(D_n) \in \{0, 1\}$, the sum of its misclassification probabilities in these two models is one. Consequently, its worst-case error is at least $1/2$.*

Proof. Take $A = \{a, b\}$, $a_0 = a$, utility equal to a binary success outcome, and no further informative observations. In both models $Y(a) \sim$ Bernoulli(0.6). In model θ_G, the archived assessment gives $q(b) = 0.4$; in θ_F, it gives $q(b) = 0.9$. Thus $\Delta_G = 0$ and $\Delta_F = 0.3$. The auditor sees only $Y(a)$ and independent noise with the same law in both models. Hence D_n has a common law P^n. Including any independent randomization in that law,

$$P_G^n(\delta = 1) + P_F^n(\delta = 0) = \mathbb{E}_{P^n}[\delta] + \mathbb{E}_{P^n}[1 - \delta] = 1. \tag{2.1}$$

One of the two nonnegative terms must be at least $1/2$. □

The unobserved alternative assessment can, for example, be in a lost briefing available to the decision maker at the time. The impossibility is epistemic: it does not imply that reading the briefing would have been useless.

Corollary 2.2 (Misfortune is not a certificate)**.** *In the construction above, both a bad outcome $Y(a) = 0$ and a good outcome $Y(a) = 1$ have positive probability under either error label. Neither observation logically certifies decision quality.*

Bad outcomes can nevertheless be statistical evidence of error when the likelihoods differ. The theorem rules out a universal inference, not all inferences. Identifiability can be restored by original decision records or by observations that distinguish models with different Δ_θ.

Illustrative archived decision	$q(a_0)$	$\max_a q(a)$	Realization
Send a reasonable invitation; receive no reply	.60	.60	0
Buy a negative-value gamble; happen to win	-.20	.00	1
Make a careless remark; nobody objects	.30	.80	1
Decline an unfavorable bet; watch it win	.00	.00	0

Table 1: Illustrative utility units, not empirical estimates. The last column records realized utility of the selected action; it does not determine the expected-utility deficit. With $\varepsilon = 0$, only the middle two rows are errors.

3 Bayesian Estimation of Having Fucked Up

Give the archived models prior weights $w_0(\theta) > 0$. Conditional on the chosen action and the observation mechanism, evidence Z_n updates the weights by

$$w_n(\theta) = \frac{w_{n-1}(\theta) p_\theta(Z_n \mid D_{n-1})}{\sum_\vartheta w_{n-1}(\vartheta) p_\vartheta(Z_n \mid D_{n-1})}, \qquad p_n = \sum_{\theta \in \mathsf{F}} w_n(\theta). \tag{3.1}$$

The posterior describes uncertainty about a fixed ex ante deficit; it does not replace original expectations with realized utilities. If declaring error when none occurred costs $c_{\mathrm{FP}} > 0$, and denying error when it occurred costs $c_{\mathrm{FN}} > 0$, the Bayes classification rule declares error precisely when

$$p_n > \tau, \qquad \tau = \frac{c_{\mathrm{FP}}}{c_{\mathrm{FP}} + c_{\mathrm{FN}}}. \tag{3.2}$$

At equality either declaration minimizes loss. Classification costs need not equal the costs of apologizing.

Proposition 3.1 (When evidence eventually settles it)**.** *Suppose Z_i are conditionally iid on a finite alphabet, all model probabilities are strictly positive, and the true model θ_* belongs to Θ. If every opposite-label model θ satisfies $D_{\mathrm{KL}}(p_{\theta_*} \| p_\theta) > 0$, then $p_n \to \mathbf{1}_{\mathsf{F}}(\theta_*)$ almost surely.*

Proof. For each opposite-label model, posterior log odds relative to the truth are

$$\log \frac{w_n(\theta)}{w_n(\theta_*)} = \log \frac{w_0(\theta)}{w_0(\theta_*)} + \sum_{i=1}^{n} \log \frac{p_\theta(Z_i)}{p_{\theta_*}(Z_i)}. \tag{3.3}$$

Dividing by n, the strong law gives limit $-D_{\mathrm{KL}}(p_{\theta_*} \| p_\theta) < 0$. Each such odds ratio tends to zero. Since $w_n(\theta) \le w_n(\theta)/w_n(\theta_*)$ and there are finitely many models, total opposite-label mass vanishes. □

Example. Let the two classes F, G each contain one model, with $p_0 = .2$. Three conditionally independent binary cues have likelihood ratios $2, 3, 1/4$ for their observed values. These are valid, for instance, using observed-value probabilities $(.8, .4)$, $(.6, .2)$, and $(.2, .8)$ under (F, G). Posterior odds are $1/4, 1/2, 3/2, 3/8$, giving probabilities $.2, 1/3, .6, 3/11$. A long silence can be followed by the discovery that the phone was off. The phone need not have read the message and formed a position.

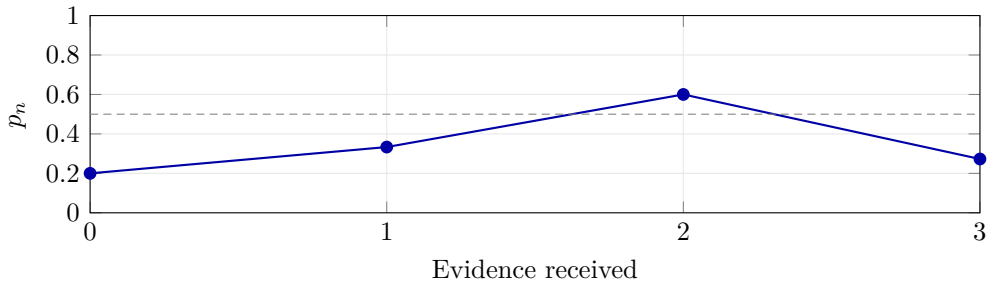


Figure 1: Exact Bayesian updates in the example. The dashed line is the equal-cost diagnostic threshold. Connecting segments guide the eye; no evidence arrives between observations.

4 Counterfactual Regret

The question "What if I had just said something normal?" requires a joint law for potential outcomes, not merely a more vivid description of the observed one. Causal models distinguish observational association, intervention distributions, and individual counterfactuals [2]. Here assume that all $Y(a)$ are defined on a common probability space and that their utility marginals agree with the archived assessment in (1.1). Fix θ, I_0 throughout this section.

Define realized hindsight regret and its ex ante expectation by

$$r = \max_{a \in A} u_0(Y(a)) - u_0(Y(a_0)), \qquad \bar{r} = \mathbb{E}[r \mid I_0]. \tag{4.1}$$

The maximum uses a hypothetical oracle that sees all outcomes. It is not an action available to the original individual.

Proposition 4.1 (The price of an imaginary oracle)**.** *With bounded utilities,*

$$\bar{r} = \Delta_\theta + H_\theta, \qquad H_\theta = \mathbb{E}[\max_a u_0(Y(a)) \mid I_0] - \max_a q_\theta(a) \geq 0. \tag{4.2}$$

Thus positive expected hindsight regret does not imply a decision error.

Proof. Take expectations in (4.1), then add and subtract $\max_a q_\theta(a)$. The expected maximum is at least the expectation of every fixed action. Maximizing these expectations proves $H_\theta \geq 0$. □

For example, if two independent actions each yield a fair binary payoff, choosing either has $\Delta = 0$, but $\bar{r} = 1/4$: precisely the probability that the selected action fails and the unselected action succeeds. A perfectly rational decision maker can therefore sustain a modest but mathematically inexhaustible supply of regret.

Theorem 4.2 (The Hindsight Impossibility Theorem)**.** *Suppose binary potential outcomes satisfy* $\mathbb{P}(Y(a) = 1) = .6$ *and* $\mathbb{P}(Y(b) = 1) = .7$*. These marginals, even if identified by randomized experiments, do not identify the probability that b would have rescued an observed failure under a. Its sharp identified set is*

$$\mathbb{P}(Y(b) = 1 \mid Y(a) = 0) \in [1/4, 1]. \tag{4.3}$$

Proof. Put $t = \mathbb{P}(Y(a) = 0, Y(b) = 1)$. The four cells, in the order $(0,1), (0,0), (1,1), (1,0)$, have probabilities

$$t, \quad .4 - t, \quad .7 - t, \quad t - .1. \tag{4.4}$$

Nonnegativity is equivalent to $.1 \leq t \leq .4$, and every value in this interval defines a valid joint law with the specified marginals. Conditional rescue probability is $t/.4$, proving both the bounds and their attainability. □

The expected benefit of choosing b over a is nevertheless identified as $.7 - .6 = .1$. One may know which action had greater expected utility without knowing whether it would have helped on this occasion. The realized oracle regret has expectation $t \in [.1, .4]$, consistent with (4.2).

Remark 4.3 (Hindsight and changing preferences)*.* Evaluating the past with $\mathbb{E}[u_0(Y(a)) \mid I_0, Y(a_0)]$ changes the information set. Evaluating it with a later utility u_n changes the objective. Either may guide a present decision, but neither is the original Δ_θ. We assume commitment to u_0 for diagnosis and a time-consistent present objective for recovery; preference reversals are outside the stopping theorem below.

5 Damage Control and the Optimal Recovery Policy

Diagnosis concerns a historical decision. Recovery concerns a current state. Let $X \in \{0,1\}$ indicate whether there is presently repairable damage, and let $b = \mathbb{P}(X = 1 \mid D_n)$. The joint posterior of (θ, X) is needed to connect b with p_n; equality is not assumed. Someone may have made a poor decision whose consequences have not arrived, or a sound decision whose consequences now require attention.

Consider a single terminal repair opportunity of intensity $v \geq 0$. In current utility units let its expected net gain over acceptance be

$$g(b, v) = bK(1 - e^{-kv}) - hv - \frac{\lambda}{2}v^2, \qquad K, k, h, \lambda > 0. \tag{5.1}$$

Here K is the maximum recoverable benefit conditional on damage, k measures repair effectiveness, h is marginal effort cost, and $\lambda v^2/2$ represents escalating intrusion or collateral harm. This is a stylized control response, not an empirically calibrated law of apologies. Remaining loss not included in K is irrecoverable in this action set.

Proposition 5.1 (The Damage Control Threshold)**.** *The maximizer $v^*(b)$ is unique. It equals zero if $bkK \leq h$; otherwise it is the unique positive solution of*

$$bkKe^{-kv^*} = h + \lambda v^*. \tag{5.2}$$

Where positive, $v^*(b)$ increases with b and decreases with h.

Proof. We have $g_v = bkKe^{-kv} - h - \lambda v$ and $g_{vv} = -bk^2Ke^{-kv} - \lambda < 0$. Also $g_v \to -\infty$. Strict concavity and the boundary derivative give the threshold and unique root. Implicit differentiation of (5.2) yields

$$\frac{\partial v^*}{\partial b} = \frac{kKe^{-kv^*}}{bk^2Ke^{-kv^*} + \lambda} > 0, \qquad \frac{\partial v^*}{\partial h} = \frac{-1}{bk^2Ke^{-kv^*} + \lambda} < 0. \tag{5.3}$$

□

Corollary 5.2 (The Double Fuck-Up Principle)**.** *If $b > 0$, increasing intervention beyond v^* strictly reduces expected utility. Moreover $g(b, v) < 0$ whenever $hv + \lambda v^2/2 > bK$. A sufficiently elaborate follow-up can therefore be worse than doing nothing, even when the original damage is certain.*

Example. With $(K, k, h, \lambda) = (4, 1, 1, 1)$ and $b = .6$, (5.2) gives $v^* \approx .4820$ and net gain $g \approx .3197$. Intensity $v = 3$ instead gives $g \approx -5.2195$. The model assigns no remedial bonus for explaining the previous explanation.

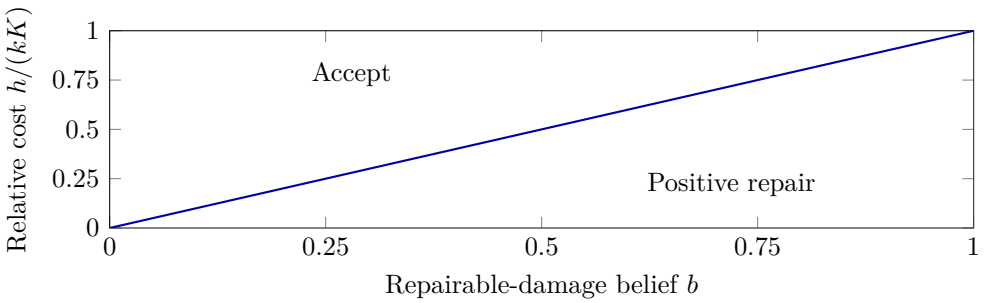


Figure 2: Terminal repair phase diagram from Proposition 5.1. On and above the diagonal, $v^* = 0$; below it, repair has positive value. At earlier dates, useful information can introduce a waiting region.

6 The Optimal Time to Give Up

For sequential recovery, use a finite-horizon partially observed control model with decision dates $t = 0, \dots, T$. Let S_t be a finite latent state containing damage, responsiveness, and any relevant historical type. Its posterior β_t is a sufficient statistic under known controlled transition and observation kernels. The record type θ, if included, remains fixed. Current utility is additive, bounded, and discounted at $0 < \gamma \leq 1$. Acceptance terminates the modeled problem with incremental payoff zero; unavoidable losses have already been included in that baseline.

At $t < T$, a finite set C_t of continuation actions includes any available repairs and information acquisitions. Action c has expected incremental reward $r_t(\beta, c)$, including its costs, and yields an updated belief β' by Bayesian filtering after state transition and observation. Standard belief-state dynamic programming [3] gives

$$V_T(\beta) = 0, \quad Q_t(\beta, c) = r_t(\beta, c) + \gamma \mathbb{E}[V_{t+1}(\beta') \mid \beta, c], \quad V_t(\beta) = \max\{0, \max_{c \in C_t} Q_t(\beta, c)\}. \tag{6.1}$$

Terminal repairs can be represented at $T - 1$; the continuous model in Section 5 is a separate terminal specialization. Equation (6.1) permits repeated interventions only when their changed state and accumulating costs are modeled.

Proposition 6.1 (Stopping criterion)**.** *An optimal policy stops at the first visited belief satisfying* $\max_c Q_t(\beta_t, c) \leq 0$*, with ties resolved toward acceptance. Otherwise it takes an action attaining the largest continuation value. Its expected discounted payoff is* $V_t(\beta_t)$*.*

Proof. At T acceptance is the only action and has value zero. Suppose the claim holds at $t + 1$. Conditional on a chosen continuation action, the induction hypothesis bounds every future policy by V_{t+1} and supplies a policy attaining it. Thus its optimal conditional value is exactly Q_t. Maximizing this against the available zero payoff proves the recursion and the stated rule. Backward induction finishes the proof. □

The comparison uses future policy value, not just immediate improvement. A myopic rule can stop too early when a modestly costly diagnostic observation would expose a highly effective repair.

For a static state, an experiment E costing c_E, and a terminal choice $a \in \mathcal{A}$ with bounded payoff $R(a, S)$, define

$$M(\beta) = \max_a \mathbb{E}_\beta R(a, S), \qquad \mathrm{EVSI}(E) = \mathbb{E}[M(\beta^E)] - M(\beta), \tag{6.2}$$

$$\mathrm{EVPI}(\beta) = \mathbb{E}_\beta[\max_a R(a, S)] - M(\beta). \tag{6.3}$$

For finite $\mathcal{A}$, the decision maker can ignore the experiment, so $0 \leq \mathrm{EVSI} \leq \mathrm{EVPI}$: the upper bound follows by bounding each conditional expected payoff by the conditional expectation of $\max_a R(a, S)$. Acquire this experiment exactly when $\mathrm{EVSI} > c_E$ (ties are indifferent), provided it causes no state change or delay loss. More informative experiments improve decision opportunities in the sense studied by Blackwell [4]; these elementary bounds suffice here.

Corollary 6.2 (The Just Go to Bed Corollary)**.** *Suppose at every remaining date the only continuation is rumination: it leaves the state and belief unchanged, produces no incremental benefit, and costs* $c_t > 0$*. Then acceptance is uniquely optimal now and at every subsequent decision date.*

Proof. Backward induction in (6.1) gives $V_t = 0$ and $Q_t = -c_t < 0$. □

This requires genuinely uninformative rumination. A new memory, an external reply, or restorative sleep is not covered by that assumption. Six hours of reprocessing the same evidence supplies six hours of cost, not six independent likelihood ratios.

7 Main Result: The Fuck-Up Theorem

The following synthesis separates a historical label, current opportunities, and the decision to continue. Its content is conditional on the specified model; it does not turn a subjective archive into an objective moral tribunal.

Theorem 7.1 (The Fuck-Up Theorem)**.** *Assume the finite archived-model specification of Sections 1–3, positive diagnostic costs, the bounded finite-horizon recovery model of Section 6, and its known controlled kernels. Given a current joint posterior β and its error marginal $p = \mathbb{P}(\mathsf{F} \mid D)$:*

(i) It is Bayes-optimal to conclude that a material decision error occurred if $p > c_{\mathrm{FP}}/(c_{\mathrm{FP}} + c_{\mathrm{FN}})$, and to conclude the opposite if the inequality reverses. Either label is optimal at equality.

(ii) Define economic recoverability *at (t, β) by $V_t(\beta) > 0$. It holds exactly when some continuation action has $Q_t(\beta, c) > 0$. On such a belief, choose a maximizing action and update; otherwise acceptance is optimal.*

(iii) Neither the diagnostic label nor p alone determines economic recoverability in general. In the terminal model (5.1), recoverability holds exactly when $bkK > h$.

(iv) For a static terminal choice, every experiment has gross value at most EVPI*. Conditionally uninformative, state-preserving rumination with positive costs and no other continuation opportunities should terminate immediately.*

Proof. The posterior losses of declaring error and denying it are respectively $c_{\mathrm{FP}}(1 - p)$ and $c_{\mathrm{FN}}p$. Comparing them proves (i). The finite-horizon backward-induction argument in Proposition 6.1 proves (ii), including attainment of the value and optimal acceptance at zero. For (iii), Proposition 5.1 gives the terminal characterization. To show that p does not suffice, fix any p and keep the same marginal distribution of θ in two joint posteriors, but set $X = 0$ almost surely in one and $X = 1$ almost surely in the other. Taking $0 < h < kK$ yields respectively zero and positive terminal value. Finally, the conditional maximum argument after (6.2) proves the information bound; Corollary 6.2 proves the last assertion. □

A numerical reversal: a worse diagnosis, a quieter phone. Use the two diagnostic types of Section 3 and suppose

$$\mathbb{P}(X = 1 \mid F) = .1, \qquad \mathbb{P}(X = 1 \mid G) = .8. \tag{7.1}$$

Assume the cues in that example are independent of X conditional on type. Interpret F as a foreseeably poor decision whose consequences are largely unresponsive to a follow-up; under G, the issue is usually a readily correctable misunderstanding. Equation (7.1) is an assumption about repairability, not a general claim that poor decisions are less repairable. Then

$$b = .1p + .8(1 - p) = .8 - .7p. \tag{7.2}$$

Let $(K, k, h, \lambda) = (4, 1, 2, 1)$. Initially $p = .2$, hence $b = .66$ and $4b = 2.64 > 2$: positive repair is optimal. After the first two cues, $p = .6$, hence $b = .38$ and $4b = 1.52 < 2$: acceptance is optimal. With equal diagnostic costs, the rational label has switched from "no" to "yes," while the rational intervention switches from positive to zero. Bad news about the past need not be a request for another message.

The belief b can also rise while p falls, and optimal action can change without the diagnostic threshold ever being crossed. In a richer model, even (p, b) may be insufficient: different beliefs about responsiveness or time remaining can produce different continuation values. The posterior over the decision-relevant state is the object to retain.

8 Conclusion

A decision is evaluated against feasible alternatives using the information and utility criterion available when it was made. An outcome is a realization of uncertainty. Their association may be informative, but their equivalence is false. In particular, an unfavorable realization does not retroactively create an attractive alternative, and a favorable realization does not certify that the chosen action was sensible.

The framework supports six distinctions that ordinary retrospective language tends to collapse. A *genuine fuck-up* has $\Delta_\theta > \varepsilon$. An *unfortunate outcome* combines $\Delta_\theta \leq \varepsilon$ with a bad realization. An *unrecognized fuck-up* combines an actual error with a current diagnostic decision to deny it, whether or not consequences have arrived. A *recoverable fuck-up* combines an actual error with positive current recovery value. An *economically irreversible fuck-up* combines an actual error with zero recovery value under the available actions. Finally, a *false alarm* is a declaration of error when $\Delta_\theta \leq \varepsilon$. These descriptions overlap: they distinguish historical quality, outcome, awareness, and repairability rather than partitioning people into six mutually exclusive species.

Economic irreversibility is deliberately weaker than physical irreversibility. A situation may be technically alterable but too costly to improve in expected utility. Conversely, $V_t > 0$ promises positive expected net benefit, not guaranteed restoration. Acceptance is optimal only relative to the action set, horizon, costs, and probability model supplied. An omitted effective repair or a misspecified likelihood can change the conclusion.

The practical procedure is consequently modest. First reconstruct the original decision without smuggling in its outcome. Then update beliefs using genuinely new evidence, keeping the possibility of non-identifiability explicit. Finally compare present interventions, information acquisition, and acceptance using their prospective value. A confession threshold and a repair threshold answer different questions. One can reasonably apologize without proving a historical error, or acknowledge an error without finding another intervention worthwhile.

No numerical probabilities in this paper are empirical estimates of social behavior. The binary states, finite model classes, risk-neutral example utilities, and saturating repair curve are transparent simplifications. Real applications require an appropriate utility scale, calibrated observation mechanisms, and an honest accounting of information and effort costs. Causal rescue claims require joint assumptions beyond marginal success rates. Sequential recommendations require stable preferences or an explicit model of their evolution.

These limitations leave the principal conclusion intact: uncertainty about what happened does not prevent a rational decision about what to do next. The past need not be acquitted before the present is allowed to improve. Sometimes one fucked up; sometimes the world merely sampled from the left tail. In neither case does the universe owe the decision maker a sufficiently long evening to make the distinction certain.

References

[1] J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Second edition. Springer, 1985. doi:10.1007/978-1-4757-4286-2.

[2] J. Pearl. *Causality: Models, Reasoning, and Inference*. Second edition. Cambridge University Press, 2009. https://bayes.cs.ucla.edu/BOOK-2K/.

[3] R. Bellman. *Dynamic Programming*. Princeton University Press, 1957.

[4] D. Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24(2):265–272, 1953. doi:10.1214/aoms/1177729032.