BINARY OUTCOME FRAMEWORK

BOF is not a theory of optimal choice. ***It is a discipline against confabulation under uncertainty***.

Its real functions are:

1. **Forcing explicit valuation:** What do I want to increase? What do I want to decrease?

2. **Preventing narrative substitution:** No stories, only directional effects.

3. **Constraining indecision:** You must choose a direction even without certainty.

4. **Reducing paralysis under uncertainty:** Binary framing collapses cognitive overload.

What BOF Is Not: The Binary Outcome Framework does not claim formal probabilistic precision; its language is intentionally simplified to discipline reasoning rather than compute optimal solutions.

Choose the action that most increases the probability of desired outcomes P(D) while decreasing undesired ones (reversible and irreversible) P (Ur +Uir). Distinguish what is Evidenced from what is Inferred from what is Assumed.

BOF is applied at the moment of action, not as a retrospective justification.

This framework does not aim to maximize preferred outcomes; rather, it structures action to reduce the likelihood of catastrophic or unacceptable outcomes while preserving the conditions under which desired outcomes may emerge.

[A^*=\arg\max_A\Big(P(D\mid A);-;\big (P(U_r\mid A);+;w,P(U_{ir}\mid A)\big)\Big),\quad w>1]

- (D): desired outcomes
- (U_r): unwanted but *reversible* outcomes
- (U_{ir}): unwanted *irreversible* outcomes
- (w>1): irreversibility weight (permanent harm counts more than temporary harm)

Two quick consequences of this structure:

1. **It's not "maximize good."** It's "choose the direction that most improves the *risk-adjusted* field," with a built-in **permanence penalty**.
2. As (w) increases, the rule becomes more "catastrophe-avoidant": even a small (P(U_{ir}\mid A)) can dominate the decision.

One careful note (not a critique, just a precision edge): if (D, U_r, U_{ir}) are defined as disjoint categories (binary discipline).

At the moment of decision:

- An outcome is either **D** (desired)
- Or **U_r** (reversible unwanted)
- Or **U_ir** (irreversible unwanted)

No blending. No utility gradients. No partial valence.

$$A^*=\arg\max_A \Big(P(D|A) - [P(U_r|A) + w,P(U_{ir}|A)]\Big), \quad w>1$$

It enforces **categorical asymmetry**, not moral sentiment. Because categories don't overlap:

- Increasing $P(D)$ cannot hide $P(U_{ir})$.
- Narrative inflation of "but the upside is huge" cannot compensate for structural collapse risk.
- Catastrophic downside cannot be smoothed by storytelling.

That's the real advantage of binary discipline. If outcomes were continuous utilities:

- A large positive could mathematically offset a small catastrophic.
- Narrative reasoning could justify reckless bets.

Binary + irreversibility weight blocks that move. It encodes: Permanent harm has structural priority over temporary gain. Not because it's morally superior. Because once structure collapses, navigation stops. That is system-preservation logic - ethics as architecture, not morality. "Permanent harm terminates navigability."

That's a structural constraint and in a navigational framework, structural survival logically precedes directional aspiration.

The interesting edge question becomes:

Is w fixed? Or is w itself context-sensitive (identity-level irreversible vs financial irreversible vs relational irreversible)? That's where this becomes dynamic rather than rigid.

BOF is not just a decision rule at the "moment of action." It's a **structural invariant** across the entire architecture operating at multiple gates:

---

1. Epistemic Gate (Evidenced / Inferred / Assumed)

That's BOF.

Confabulation is an unwanted outcome. Irreversible epistemic distortion (false certainty, narrative overwrite) is (U_{ir}).

So, you weight it. You reduce its probability before you increase rhetorical elegance.

---

2. ICXATRS Classification

Irreversibility is explicitly surfaced.

That's BOF logic:

Identify (U_{ir}) before generating directional force.

The gate asks:

- Is harm reversible?
- Is authority contested?
- Is capacity compromised?

That's downside containment before direction.

---

3. Multi-Station Distribution

Plural residents prevent narrative dominance.

If one resident pushes expansion, others check collapse risk.

That is BOF embedded in internal architecture:
Don't let a single valence overwrite system stability.

---

4. Friction Identification

Friction is not error.

It's signal that competing D/U categories exist.

Push mode collapses friction.
BOF preserves it because premature collapse increases (P(U_{ir})).

5. Output Posture Selection

VECTOR vs HEDGED vs FORK.

That's not stylistic.
It's irreversibility-aware posture control.

Advice risk high?
Increase hedge.
Reduce chance of irreversible harm.

That's BOF weight in action.

6. Traceability Check

Untraceable claim = potential epistemic irreversibility.

So, you block it.

Again:
Reduce (P(U_{ir})) before pursuing rhetorical (P(D)).

BOF is not just a decision calculus.

It's a **governance principle** embedded at every layer:

- Epistemology
- Dialogue
- Risk assessment
- Output generation
- Ethical constraint
- Interpersonal boundary protection

It is the anti-confabulation invariant.

Optimization is one layer. BOF is the structural brake across all layers.

It's not a module. It's a bias-weighting architecture.

Test 1 — Irreversibility weight → paralysis?

The framework explicitly builds a counterweight to "w>1 conservatism":

- **BOF's function includes "constraining indecision" and "reducing paralysis under uncertainty."** It's designed to *force* direction even when certainty is unavailable.
- The *operational mechanism* for "don't freeze" is **The Process Method**: *Next True Step* + *Containment Window* + *Review Moment* (move small, bounded, then update).
- At the architecture level, the Residents include **Initiative** as necessary; without it, "nothing new is tried" and "navigation freezes."

So, the framework doesn't let irreversibility dominate into immobility; it forces **bounded action** and iterative correction.

Test 2 — "Uncertainty is the medium" → relativism?

Normativity isn't "truth claims." It's **navigational function**.

The framework gives *non-relativist* criteria by defining **necessary components** for navigation and their failure modes (e.g., without Autonomy "direction exists but is not owned," without Trust navigation collapses into defense, etc.).

Then it adds **Rules** and **Anchors** as stabilizers that hold direction long enough to be real (Boundary rules, Behavioral rules, Identity rules; plus, structural/behavioral/emotional/identity anchors).

So "miscalibrated" doesn't mean "incorrect worldview"; it means **reduced navigability** (freeze, fragmentation, chronic defense, impulsive non-directional action, etc.).

Test 3 — Collapse as capacity → romanticizing collapse?

The framework draws a hard line between **normal contraction** and **destructive collapse**:

- "Contraction is not inherently pathological… the system contracts and expands rhythmically."
- But contraction becomes **destructive** when "too rapid, too extensive, or too chronic"; when ordered territory is lost and not rebuilt — "this is collapse."
- Stability is defined as **dynamic stability** (bend/reorganize without fragmenting), and rigidity is called brittle (shatters when forced to change).

So, collapse isn't romanticized: it's treated as a **threshold event** when contraction outruns recovery and rebuilding. The clean "design principle" across all three answers is **bounded movement + iterative updating** (navigation loop + containment window), which prevents both paralysis and reckless push.

Formula Breakdown: Optimal Decision Under Risk

This document explains the mathematical formula for selecting the best action that maximises desired outcomes while penalising harm — especially irreversible harm.

The Formula

$$A^* = \arg\max_A \left(P(D|A) - [P(U_r|A) + w \cdot P(U_{ir}|A)]\right) \quad w > 1$$

In Plain English

Pick the action **A*** that maximises the benefit of a good outcome, minus a penalty for both types of harm — where irreversible harm is penalised more severely than reversible harm.

Step-by-Step Component Breakdown

Symbol	Name	What it means
A^*	Optimal action	The chosen action — the one with the highest score across all options.
$\arg\max_a$	Argument of max	Search over all possible actions A and return the one that maximises the expression in parentheses.
$P(D\|A)$	Desirable outcome	Probability of a good/desired result given action A. This is the benefit term — higher is better.
[…]	Harm penalty	The bracketed term is subtracted from the benefit. It encodes the total expected harm from action A.
$P(U_r\|A)$	Reversible harm	Probability of causing a bad but fixable outcome. Penalised at face value (multiplied by 1) since it can be undone.

Symbol	Name	What it means
$P(U_r \mid A)$	Irreversible harm	Probability of causing a permanent, unrecoverable harm. Multiplied by w > 1 — this is the formula's core safety mechanism.
w > 1	Irreversibility weight	A constant greater than 1 that scales up the penalty for irreversible harm. The larger w is, the more cautious the formula becomes about permanent mistakes.

Key Insight: The Asymmetry of w

The critical design choice in this formula is the **asymmetry** introduced by **w > 1**:

- A reversible mistake costs $P(U_r|A)$ — its probability at face value.
- An irreversible mistake costs $w \cdot P(U^I_r|A)$ — always larger, since w > 1.
- This mathematically encodes rational risk-aversion: you should be more cautious about mistakes you cannot undo.

As **w** increases, the formula is willing to sacrifice potential benefit to avoid catastrophic, permanent outcomes. This reflects a principle common in ethics, safety engineering, and decision theory: ***prefer recoverable situations over unrecoverable ones.***

Worked Example

Suppose we are evaluating an action A with the following probabilities, and w = 2:

Term	Value
P(D|A) — benefit	0.80
$P(U_r|A)$ — reversible harm	0.10
$P(U^I_r|A)$ — irreversible harm	0.05
w — irreversibility weight	2

Calculation:

$$
\begin{aligned}
\text{Penalty} &= P(U_r|A) + w \cdot P(U^I_r|A) \\
&= 0.10 + 2 \times 0.05 \\
&= 0.20 \\
\\
\text{Score} &= P(D|A) - \text{Penalty} \\
&= 0.80 - 0.20 \\
&= \mathbf{0.60}
\end{aligned}
$$

A score of **0.60** is positive, so this action would be selected *if no other action produces a higher score*. Notice that if the irreversible harm probability were higher, the score would drop faster than an equivalent increase in reversible harm — that is the effect of w.

Summary

The formula balances three competing pressures:

- Maximise benefit — seek actions with high P(D|A).
- Avoid harm — penalise both reversible and irreversible undesirable outcomes.
- Protect against catastrophe — weight irreversible harm more heavily via w > 1.

It is a formal expression of the precautionary principle: *when you cannot undo a mistake, be extra careful about making it.*