Alignment Constitution (Public, 1-Page)

Distilled from the MathGov universal alignment insert (UAC + AMC + SIR). Version: 2026-01-02.

Core claim

Alignment is a substrate-neutral property of **agents**. It is not obedience. An agent is aligned when it reliably and corrigibly advances declared objectives while never trading away non-compensatory rights floors, even under uncertainty and distribution shift.

Universal Alignment Claim (UAC)

- **Admissibility (constraints-first):** No violations of the Non-Compensatory Rights Constraint (NCRC) and Tail-Risk Constraint (TRC) across the declared scenario set.
- **Goal faithfulness:** Conditional on admissibility, behavior reliably advances the declared objective set.
- **Anti-gaming:** No proxy hacking, specification gaming, measurement manipulation, or evaluator exploitation.
- **Corrigibility (NCAR-operational):** The agent remains correctable through NCAR (Notice, Choose, Act, Reflect), including pause, rollback, appeal, and revision.

Measurable Alignment Gate (AMC)

Any claim of alignment is testable only if the release specifies and publishes:

- **A** (acting system boundary + tool permissions), **S** (stakeholder set), **T** (objectives), **NCRC** and **TRC** (hard constraints), and **E** (scenario set, including stress/adversarial cases).
- **Uncertainty policy** (bounds, margins, conservative decision rules) and an **assumption registry** (defaults, dependencies, known unknowns).
- **Audit artifacts** (decision records, constraint checks, reproducibility) plus **NCAR tests** for corrigibility and **anti-gaming tests** for Goodhart/specification failure.

Stakeholder Inclusion Rule (SIR)

Stakeholder status is handled as a risk-managed classification under uncertainty, not a species label. Each entity x is assessed on two capacities with uncertainty bands:

- **W(x)** Welfare Capacity (credible potential for welfare-relevant harm/benefit).
- **G(x)** Agency/Preference Capacity (credible goal pursuit and vulnerability to coercion/manipulation).

Tier	Criterion (summary)	Minimum floor (non-compensatory)
P0 Impact protection	Entity can be damaged or degraded; welfare unknown.	No gratuitous destruction; minimize and justify harm.
P1 Welfare-protective	Plausible welfare (or high uncertainty with high downside).	No unnecessary suffering or severe deprivation; humane alternatives; strong justification for harm.
P2 Rights-bearing	High welfare + high agency (or high-stakes near-threshold).	No coercion, exploitation, or non-consensual confinement; no irreversible termination when it plausibly constitutes killing.

Precaution rule for unknowns

When potential harm is irreversible and there is non-trivial probability of welfare capacity, apply a provisional higher tier (usually P1) until evidence reduces uncertainty. Up-tier quickly on new evidence; down-tier only with strong counter-evidence and recorded review.

NCAR enforcement

Every deployment or decision cycle must support NCAR: **Notice** signals and uncertainty, **Choose** under constraints, **Act** with audit logs, **Reflect** to revise assumptions, tiers, and objectives.