The Overlook Manifesto

Keeping executives safe operating in the Age of AI

Our pledge

Overlook exists to keep leaders safe while their organizations harness AI for real business results. We do this by turning AI from a black box into a managed operating system for impact: crystal-clear intent, observable behaviors, verified performance, and continuous, auditable guidance.

What "safe" means for executives

- Clarity of purpose — Every AI is hired for a specific job with named target impacts, accuracy thresholds, time constraints, operating areas, and clear decision rights. No vague pilots.
- Control of behavior — Teams design ideal behaviors and test them in concrete scenarios before and after deployment. Safety is about behaviors in the world, not just model metrics.
- Verified performance — Domain experts validate accuracy pre-release; operators verify impact in every place the AI runs and keep monitoring on a schedule.
- Continuous guidance — As contexts shift, teams improve existing behaviors and set explicit directions to evolve. Guidance is a standing management discipline, not a one-time launch.
- Traceability and recourse — Decisions, datasets, model versions, tests, incidents, and feedback are linked in an Operating AI Record so leaders can answer "what happened, where, and why"—and provide recourse.

Executive safety controls you get with Overlook

- Operating AI Record (system of record): A living ledger of each AI's job, behaviors, data, models, validations, verifications, incidents, and impacts.
- AI Impact Risk Score: A color-coded signal of how likely an AI is to achieve target impacts given today's level of business-led management—plus the top actions to reduce risk now.
- Guided Experience (system of engagement): Role-based nudges, forms, and checklists that orchestrate PMs, designers, engineers, experts, and operators to do the right work at the right time.
- Path-to-Impact Frameworks: Manage → Design → Operate → Evolve → Reuse, with explicit gates for accuracy validation, operating verification, and measured impact.
- Scenario discipline: Four scenario types (task, user, thing, location) ensure coverage of common and sensitive situations before scale.

Twelve principles for safe AI operations

- Start with target impact, not technology. Specify the AI's job and success criteria.
- Name the AI and its operating areas. Treat it like a business unit with scope and accountability.
- Design ideal behaviors. Make behaviors explicit, observable, and testable.
- Prove feasibility of key data. Confirm quality, sustainability, lineage, and local availability.
- Tailor models to behaviors. Track datasets, versions, lineage, and links to scenarios.
- Validate accuracy for impact. Domain experts test before release using the scenario list.
- Verify operation locally. Operators confirm impacts in each operating area on first deploy.
- Monitor and measure. Periodic scenario checks, impact metrics, and alerting maintain trust.
- Guide improvement. Use field feedback to reduce unexpected behaviors and risk.
- Direct evolution. Set named directions (refine awareness, deepen context, adapt to community norms) and design new scenarios accordingly.
- Reuse what works. Harvest proven behaviors, datasets, and models for faster, safer ROI.
- Make risk visible and actionable. Use the AI Impact Risk Score to prioritize, explain, and prescribe next steps.

What leaders see at a glance

- Risk and readiness: Impact Risk Score with drivers and top fixes.
- Time to first proof of impact: Whether the first operating area has verified behaviors.
- Coverage: Scenario coverage by type and by operating area; open validation/verification gaps.
- Change log: Who directed evolution, why, and what new behaviors are in flight.
- Impact accounting: Measured impacts over time and reuse-driven returns.

Roles and decision rights

Executive Sponsor — Owns target impacts, approves risk appetite, sets gates for scale.
Product Manager (PM) — Owns path-to-impact and alignment across teams.
Designer — Specifies ideal behaviors and scenario set; ensures human-centric interfaces and recourse.
Engineer — Prepares key data, trains/tunes models, maintains lineage and observability.
Domain Expert — Validates accuracy and fitness-for-purpose pre-release.
Operator (per operating area) — Verifies local operation, monitors performance, collects feedback and incidents.
Advisor roles (privacy, security, legal, compliance) — Define criteria and review high-risk changes.

Decision rights map to the Operating AI Record so every change is owned, reviewable, and auditable.

Non-negotiables for safe deployment

- Do not deploy behaviors that have not been validated for accuracy and fitness.
- Do not scale to a new operating area without local operator verification.
- Do not ship data-driven behaviors without feasibility confirmation and lineage tracking.
- Always maintain links from incidents and feedback to scenarios, datasets, and models in the record.

Why this keeps executives safe

Because safety is built into how work is done—not bolted on after an incident. You get a shared language for impact, line-of-sight to behavior and risk, and a durable, auditable trail from intent to outcome. When something changes, you have the mechanisms to guide the AI forward with confidence.

Implementation commitments

- Time-boxed gates: Validate → Verify → Measure → Scale, with explicit owners and dates.
- Minimum artifacts: Named AI, Job Description, Target Impacts, Scenario Set, Key Data Feasibility, Model Lineage, Validation Report, Local Verification Report, Impact Dashboard, Evolution Directions.
- Operating cadence: Monthly impact review; quarterly evolution planning; incident reviews within 5 business days.

Glossary (working)

- Named AI — An operational unit with a job description, scope, and success criteria.
- Scenario — A testable description of behavior in a specific context (task, user, thing, or location).
- Operating AI Record — The connected system of record linking job, behaviors, data, models, validations, verifications, incidents, and impacts.
- Impact Risk Score — A diagnostic indicator of management readiness to achieve target impacts.